ANÁLISE DOS DADOS DE INQUÉRITOS SOCIOLÓGICOS: ESTATÍSTICAS UNIVARIADA, BIVARIADA E MULTIVARIADA

Ana Lúcia Teixeira

Passaremos, agora, a apresentar alguns exemplos de instrumentos de análise de dados, recorrendo a dados obtidos através de inquéritos sociológicos na área da violência doméstica e de género. O objectivo desta secção é o de ilustrar, com dados reais, a forma como os diferentes instrumentos de análise podem ser utilizados, para que são usados e quais os problemas mais comuns com que um/a investigador/a se depara no decurso do processo de investigação. Sem pretensões de exaustividade, procuraremos dar conta das potencialidades e limitações das técnicas de análise de dados mais recorrentemente mobilizadas na área das Ciências Sociais e, em particular, na área da Sociologia.

Nesse sentido, e com recurso ao *software* de análise estatística SPSS, apresentaremos uma variedade de técnicas, que poderão ser úteis aos/às estudantes e investigadores/as, ilustradas por dados reais. Para tal, dividiremos este capítulo em três secções: a primeira debruçar-se-á sobre estatísticas univariadas, essencialmente descritivas; na segunda, onde olharemos para os dados de uma perspectiva bivariada, serão incluídas técnicas de análise, tanto descritiva, como inferencial; por último, daremos conta de algumas das técnicas de análise multivariada mais comuns nesta área. Para a ilustração de cada caso referido, recorreremos aos *outputs* gerados pelo SPSS, para que o/a leitor/a melhor consiga acompanhar a sua interpretação.

Análise de dados univariada

A análise de dados univariada reveste-se de uma enorme importância, no contexto não apenas da exploração inicial de uma base de dados, como também da validação dos dados e da sua preparação para análises mais complexas. Os diferentes instrumentos de estatística univariada permitem-nos, num primeiro (e fundamental)

momento, a verificação de todas as variáveis, nomeadamente se existem ou não dados mal introduzidos (quando estamos a falar de bases de dados cuja construção é manual). Assim, a primeira abordagem estatística a uma base de dados deve consistir na produção de apuramentos individuais para cada uma das variáveis que a constituem. Esta tarefa permite-nos, por um lado, perceber se haverá alguma informação que tenha de ser corrigida (recorrendo à consulta do suporte no qual a informação foi recolhida) e, por outro lado, ter um primeiro contacto com a estrutura das respostas recolhidas.

Vejamos um exemplo genérico. Pedimos uma tabela da frequências da variável «estado civil», ainda no contexto de validação dos dados introduzidos. Como se observa pelo *output* gerado (Quadro 1), encontramos uma observação que, apresentando um valor fora do leque das respostas possíveis (código 12, quando os possíveis são os códigos 1, 2, 3 4 e 99), deverá ter resultado de um erro de inserção na base de dados.

Quadro 1. Tabela de frequências da variável «estado civil» na fase de exploração/validação da base de dados

v103a estado civil actual					
		Frequency	**Percent**	**Valid Percent**	**Cumulative Percent**
Valid	1 solteira	253	25,3	25,3	25,3
	2 casada/união de facto	575	57,5	57,5	82,8
	3 divorciada/separada	41	4,1	4,1	86,9
	4 viúva	127	12,7	12,7	99,6
	12	1	,1	,1	99,7
	99 ns/nr	3	,3	,3	100,0
	Total	**1000**	**100,0**	**100,0**	

Neste caso, o procedimento adequado para a correcção desta situação deverá ser identificar o questionário associado a esta resposta e verificar, no suporte original (questionário em papel, se for esse o caso), qual a resposta dada pela pessoa inquirida, procedendo depois à correcção da informação na base de dados. Mas observá-

mos ainda que três pessoas não responderam a esta questão (código 99). Neste caso, não se trata de informação mal introduzida, mas de uma resposta que não deverá ser utilizada em análises posteriores (a menos que as não respostas sejam, de facto, um item de interesse). O que é habitual fazer, neste caso, é dar a indicação de que este código, ainda que permaneça na base de dados, não deve ser considerado para os cálculos. Para tal, dever-se-á definir o código 99 como *missing value*. Assim, e após proceder a estas alterações, deveremos pedir uma nova tabela de frequências, para confirmar que as alterações foram feitas correctamente (Quadro 2).

Quadro 2. Tabela de frequências da variável «estado civil» na fase de exploração/validação da base de dados após alterações

v103a estado civil actual					
		Frequency	**Percent**	**Valid Percent**	**Cumulative Percent**
Valid	1 solteira	254	25,4	25,5	25,5
	2 casada/união de facto	575	57,5	57,7	83,1
	3 divorciada/separada	41	4,1	4,1	87,3
	4 viúva	127	12,7	12,7	100,0
	Total	997	99,7	100,0	
Missing	99 ns/nr	3	,3		
Total		**1000**	**100,0**		

Assim, podemos, agora, ter uma primeira visão da correcta distribuição desta variável.

No caso de tratar-se de uma variável métrica (ou quantitativa), o procedimento é um pouco diferente, uma vez que não faz sentido aplicar a este tipo de variáveis o mesmo procedimento utilizado para as variáveis categóricas (tanto nominais como ordinais). Em lugar de fazer tabelas de frequências, deveremos apurar estatísticas descritivas (ou até representações gráficas), que podem incluir uma grande diversidade de medidas, como

sejam a média, o desvio padrão, o mínimo, o máximo, os quartis (ou os percentis, se quisermos uma análise mais detalhada), entre outras. Também aqui o objectivo é o de corrigir eventuais erros de inserção e o de familiarização com os dados a analisar. Vejamos o exemplo da variável idade, para a qual pedimos algumas estatísticas descritivas (Quadro 3).

Quadro 3. Estatísticas descritivas da variável «idade» na fase de exploração/validação da base de dados

Descriptives				
			Statistic	**Std. Error**
v102 idade	Mean		44,06	1,086
	95% Confidence Interval for Mean	Lower Bound	41,93	
		Upper Bound	46,20	
	5% Trimmed Mean		42,84	
	Median		43,00	
	Variance		1180,140	
	Std. Deviation		34,353	
	Minimum		18	
	Maximum		999	
	Range		981	
	Interquartile Range		29	
	Skewness		21,545	,077
	Kurtosis		598,360	,155

Numa leitura rápida dos resultados, percebemos que há um erro a ser corrigido. O valor máximo encontrado é de 999, o que, no contexto desta variável, não faz sentido. É, portanto, necessário identificar o(s) caso(s) problemático(s) e confirmar, através da consulta do(s) questionário(s), se se trata de um erro de inserção ou de uma não resposta que não foi identificada como tal na estrutura da base de dados, e proceder à sua correcção.

Um outro objectivo desta primeira exploração da base de dados é o de uma primeira avaliação da necessidade de fazer recodificações nas variáveis. Sem prejuízo da possibilidade de proceder a futuros reagrupamentos e rectificações (que, na verdade, acompanham, de forma transversal, todo o processo de análise de dados), este é o momento no qual importa verificar se existem categorias que, com poucas ou nenhumas observações, se venham a revelar inúteis mais adiante. Vejamos o exemplo seguinte, relativo à distribuição do nível de instrução de 1000 inquiridos/as.

Quadro 4. Tabela de frequências da variável «nível de instrução» na fase de exploração/validação da base de dados

v106 nível de instrução					
		Frequency	**Percent**	**Valid Percent**	**Cumulative Percent**
Valid	1 Não sabe ler e/ou escrever	88	8,8	8,8	8,8
	2 Primário	379	37,9	37,9	46,7
	3 Preparatório	99	9,9	9,9	56,6
	4 Secundário	280	28,0	28,0	84,6
	5 Licenciatura	149	14,9	14,9	99,5
	6 Mestrado	3	,3	,3	99,8
	7 Doutoramento	2	,2	,2	100,0
	Total	**1000**	**100,0**	**100,0**	

Observa-se (Quadro 4) que o número de pessoas com mestrado e com doutoramento é muito reduzido, pelo que, para futuras análises, talvez faça sentido agrupar as categorias 5, 6 e 7 numa única, que poderá passar a designar-se «superior». Como é evidente, este tipo de decisões depende, em absoluto, dos objectivos do estudo. As sugestões que aqui apresentamos constituem-se apenas como linhas gerais de procedimento, que deverão ser avaliadas caso a caso.

Em suma, este diagnóstico deve ser aplicado a todas as variáveis da base de dados, antes de prosseguir para a análise da informação. Daremos, agora, alguns exemplos de apuramentos e estatísticas descritivas univariadas, tanto para variáveis de tipo categórico como métrico, que cumprem o propósito de obter uma visão global da informação recolhida e de preparar as análises estatísticas posteriores. Os *outputs* apresentados dizem respeito aos dados recolhidos através do inquérito sociológico nacional realizado no âmbito do projecto de investigação «Violência contra as mulheres», aplicado, em 1995, a uma amostra 1000 mulheres com 18 ou mais anos de idade, com uma margem de erro de 5%, para um nível de confiança de 95% (Lourenço, Lisboa & Pais, 1997).

No caso das variáveis categóricas, podemos, como vimos, extrair tabelas de frequências. Vejamos a variável «estado civil»:

Quadro 5. Distribuição da variável «estado civil»

v103a estado civil actual					
		Frequency	**Percent**	**Valid Percent**	**Cumulative Percent**
Valid	1 solteira	255	25,5	25,5	25,5
	2 casada/união de facto	576	57,6	57,6	83,1
	3 divorciada/separada	42	4,2	4,2	87,3
	4 viúva	127	12,7	12,7	100,0
	Total	**1000**	**100,0**	**100,0**	

Fonte dos dados: *Violência contra as mulheres*, SociNova (1995).

Dando conta de informação ainda muito simples, este apuramento (Quadro 5) permite-nos perceber a distribuição da população, ou, neste caso, da amostra, relativamente a esta variável. Podemos perceber que a maior parte das inquiridas (57,6%) é casada ou está em união de facto, e que o conjunto das mulheres divorciadas ou separadas é o menos representativo (4,2%).

No caso das variáveis métricas, encontramos disponível um vasto leque de estatísticas que nos permitem conhecer a distribuição destas variáveis em pormenor. Como vimos anteriormente, podemos calcular medidas de tendência central (média aritmética, média aparada a 5%, mediana, percentis e quartis), medidas de dispersão (variância, desvio-padrão, amplitude da distribuição, amplitude interquartílica), medidas de assimetria e achatamento (enviesamento e curtose).

Quadro 6. Estatísticas descritivas da variável «idade»

Descriptives				
			Statistic	**Std. Error**
v102 idade	Mean		43,09	,516
	95% Confidence Interval for Mean	Lower Bound	42,08	
		Upper Bound	44,11	
	5% Trimmed Mean		42,80	
	Median		42,50	
	Variance		266,610	
	Std. Deviation		16,328	
	Minimum		18	
	Maximum		75	
	Range		57	
	Interquartile Range		29	
	Skewness		,191	,077
	Kurtosis		-1,129	,155

Fonte dos dados: *Violência contra as mulheres*, SociNova (1995).

Como se observa (Quadro 6), a média de idades ronda os 43,09 anos, com um desvio padrão de 16,33 anos. Uma vez que o desvio padrão é calculado a partir dos desvios à média, o seu valor está dependente da magnitude desta estatística, pelo que não podemos afirmar, pela simples leitura do valor do desvio padrão, se a dispersão é elevada ou não (por exemplo, um desvio padrão de 5

pode representar uma dispersão muito reduzida numa distribuição e muito elevada noutra – depende do valor da média). Para avaliar a heterogeneidade da distribuição, poderemos calcular uma estatística adicional, que não é produzida pelo SPSS, mas que é muito útil para este propósito, e ainda para comparar duas distribuições que tenham médias diferentes. Tratando-se de uma medida de dispersão relativa, o coeficiente de variação fornece uma medida estandardizada, que varia entre 0 e 100%: quanto mais próximo de 100, mais elevada é a dispersão relativa, pelo que pode considerar-se a distribuição muito heterogénea, onde a média é pouco representativa da configuração da distribuição; pelo contrário, quanto mais próximo de 0, menor a dispersão relativa, considerando-se, assim, que essa será uma distribuição mais homogénea, e onde a média é mais representativa da estrutura da distribuição. O coeficiente de variação é dado por $CV=\left(\frac{S}{\bar{x}}\right)\cdot 100$, onde S corresponde ao desvio-padrão e $\bar{x}$ à média da distribuição. Para ilustrar a nota feita atrás, onde se disse que o mesmo valor de desvio padrão pode remeter para configurações de dispersão muito diferentes, tomemos o seguinte exercício: duas distribuições (*a* e *b*) com o mesmo desvio padrão $(DP_a=DP_b=5)$ mas com médias diferentes $(\bar{x}_a=300;\ \bar{x}_b=35)$. No primeiro caso, a dispersão relativa é dada por $CV_a=(5/300)\cdot 100=1{,}67\%$ e no segundo caso por $CV_b=(5/35)\cdot 100=14{,}29\%$. Percebemos, então, que a distribuição *a* apresenta uma dispersão relativa muito reduzida (1,67%) face à dispersão relativa da distribuição *b* (14,29%), revelando a inadequação da avaliação da dispersão de uma distribuição por via da leitura absoluta do desvio-padrão.

Uma outra medida importante para conhecer a distribuição de uma variável (ordinal ou métrica) são os percentis, que são medidas que dividem a distribuição ordenada em 100 partes de dimensão aproximadamente igual.

Quadro 7. Percentis da variável «idade»

		Percentiles						
		5	**10**	**25**	**50**	**75**	**90**	**95**
Weighted Average (Definition 1)	v102 idade	20,00	22,00	28,00	42,50	56,75	67,00	70,00
Tukey's Hinges	v102 idade			28,00	42,50	56,50		

Fonte dos dados: *Violência contra as mulheres*, SociNova (1995).

Como se observa no quadro anterior (Quadro 7), percebemos, por exemplo, que 25% das mulheres da amostra têm menos de 28 anos. Da mesma forma, 90% têm menos de 67 anos. Os percentis 25, 50 e 75 correspondem aos quartis – medidas que dividem a distribuição em quatro partes aproximadamente de igual dimensão. O percentil 50, que corresponde ao 2.º quartil, equivale ainda à mediana da distribuição.

Ao nível das estatísticas preparatórias de explorações posteriores, é ainda importante destacar os testes de aderência de Kolmogorov-Smirnov e de Shapiro-Wilk à distribuição normal. Ambos os testes são aplicados quando pretende verificar-se se a variável (métrica) segue uma distribuição normal. O teste de Shapiro-Wilk é mais robusto para amostras pequenas ($n \leq 50$), ao passo que, no caso de amostras de grande dimensão, o teste de Kolmogorov-Smirnov é o mais adequado. As hipóteses em teste são, em todo o caso, iguais para ambos os testes (H_0: a variável segue uma distribuição normal na população; H_a: a variável não segue uma distribuição normal na população), bem como a regra de decisão (rejeitar H_0 quando *p-value* < 0,05, para um nível de confiança de 95%).

Como se observa na figura seguinte (para um n=1000, devemos analisar o teste de Kolmogorov-Smirnov[1]), com um *p-value* de aproximadamente 0, rejeitamos a hipótese nula em teste, pelo que podemos afirmar que não existem evidências estatísticas para afirmar que a idade siga uma distribuição normal na população.

Quadro 8. Testes à normalidade da distribuição da variável «idade»

Tests of Normality						
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
v102 idade	,080	1000	,000	,953	1000	,000

[a]. Lilliefors Significance Correction

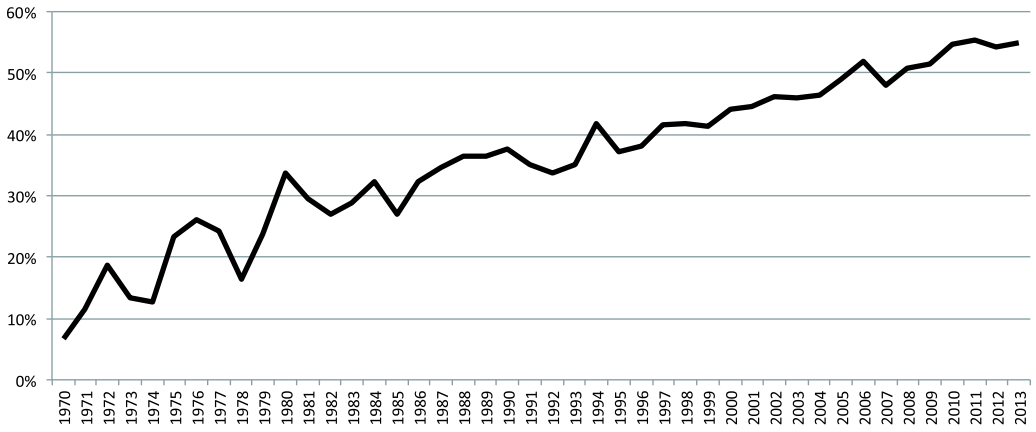
Fonte dos dados: *Violência contra as mulheres*, SociNova (1995).

1 O SPSS apresenta o teste de Kolmogorov-Smirnov com a correcção de Lilliefors, procedimento aplicável quando não são conhecidos os parâmetros populacionais (Lilliefors, 1967).

Um último tópico a abordar prende-se com a representação gráfica das distribuições. Estes instrumentos de síntese têm como principal objectivo uma comunicação mais eficaz e uma visualização mais imediata dos resultados obtidos. Servindo o propósito de pôr em evidência as ordens de grandeza e/ou a evolução dos fenómenos em observação, não devem servir de base à análise final dos dados, que deve ser feita a partir dos resultados tabelados. Existe uma multiplicidade de representações gráficas disponíveis, mas a selecção da mais adequada deve prender-se com a natureza da informação a visualizar. Vejamos alguns exemplos.

Nos gráficos de linhas, é enfatizada uma tendência, dada pela inclinação entre dois pontos. Assim, estes gráficos devem ser usados sobretudo para representar a evolução de uma variável ao longo do tempo.

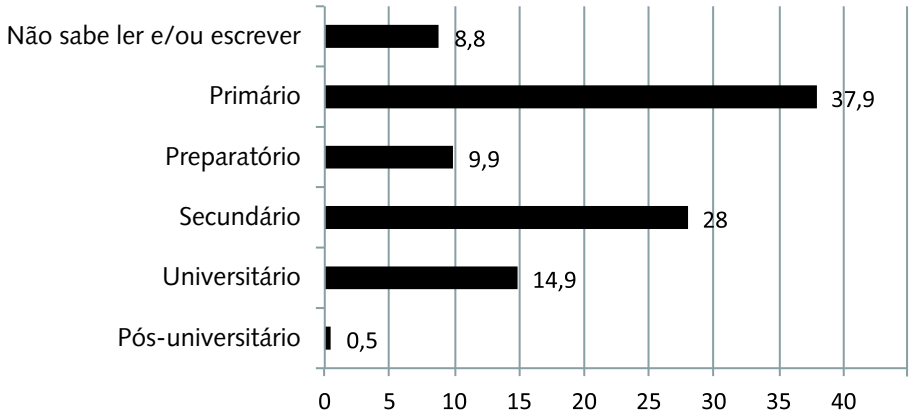
Figura 1. Evolução da proporção de doutoramentos concluídos por mulheres relativamente ao número total de doutoramentos terminados, 1970-2013 (%)



Fonte dos dados: Pordata.

Os gráficos de barras ou de colunas permitem a visualização das quantidades associadas às categorias de uma determinada variável através de rectângulos cuja altura/comprimento é proporcional ao valor representado. Também podem ser usados para representar a evolução de um fenómeno ao longo do tempo (neste caso, preferencialmente gráficos de colunas).

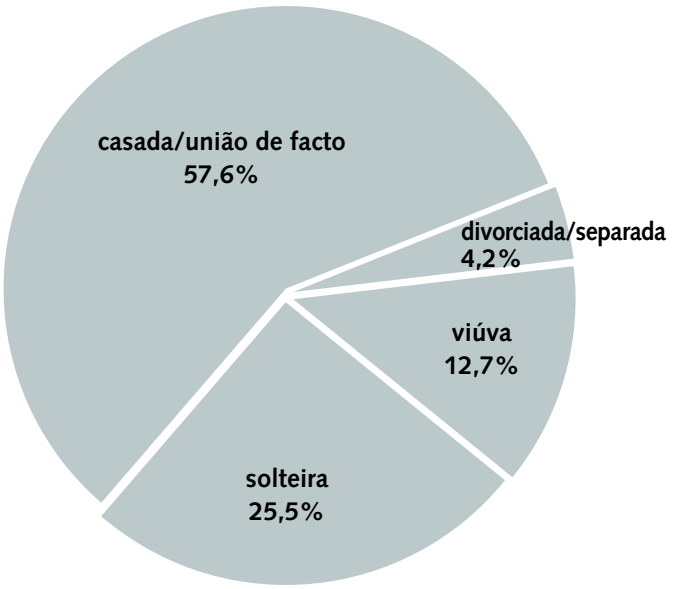
Figura 2. Nível de instrução das mulheres inquiridas (%)



Fonte dos dados: *Violência contra as mulheres*, SociNova (1995).

Os gráficos circulares consistem na representação gráfica dos resultados num círculo, dividido em sectores, cuja dimensão é proporcional ao valor da categoria representada. São indicados para quando pretende dar-se a noção do peso de cada parte relativamente ao todo, seja em termos absolutos ou relativos, num tempo preciso.

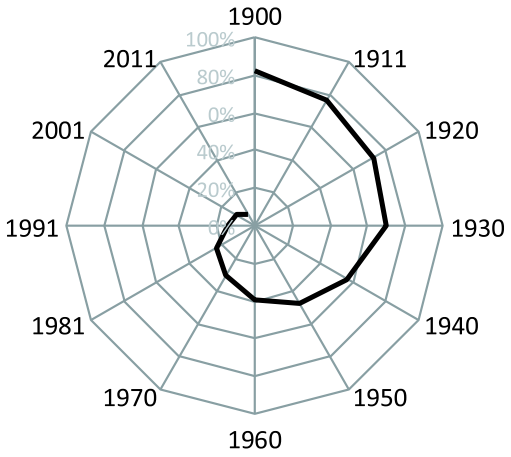
Figura 3. Estado civil das mulheres inquiridas



Fonte dos dados: *Violência contra as mulheres*, SociNova (1995).

Os gráficos polares são sobretudo utilizados para representar séries temporais ou para dar a noção da disparidade entre as várias categorias. Os dados estatísticos são apresentados através de um polígono desenhado nos «círculos» definidos pela amplitude da escala que pretendemos utilizar. No exemplo seguinte (Figura 4), observamos a evolução da taxa de analfabetismo das mulheres portuguesas no século XX, através de dados censitários. Percebe-se que partimos, em 1900, de uma taxa de analfabetismo a rondar os 80%, e que chegámos a menos de 10% em 2011. Mais uma vez, chamamos a atenção para que a análise dos dados não deve ser iniciada das representações gráficas, mas sim das estatísticas originais. Em todo o caso, percebe-se com clareza a evolução desta variável ao longo do tempo.

Figura 4. Taxa de analfabetismo das mulheres portuguesas, 1900-2011

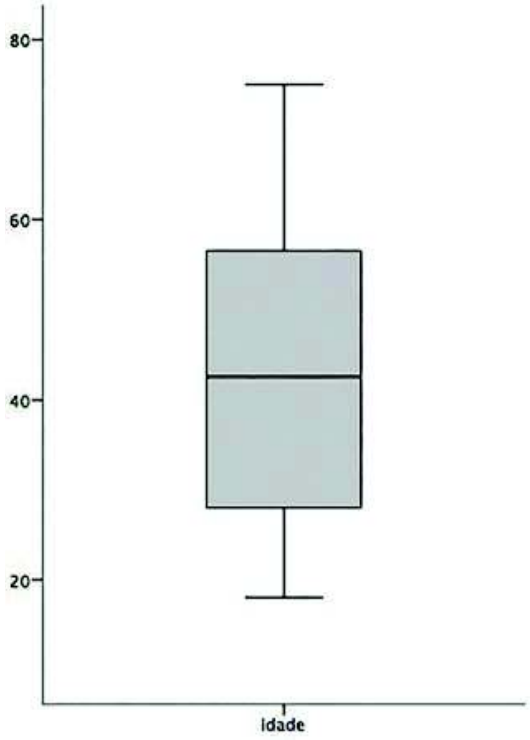


Fonte dos dados: INE.

O gráfico de extremos e quartis (também designado por caixa de bigodes ou por *boxplot*) representa graficamente a distribuição de uma variável (ordinal ou métrica) a partir dos quartis, mínimo e máximo. Apresenta ainda os *outliers* (observações com valores 1,5 vezes superiores à soma dos valores do 3.º quartil e da amplitude interquartílica; ou 1,5 vezes inferiores ao valor do 1.º quartil menos a amplitude interquartílica) e os valores extremos

(observações com valores três vezes superiores à soma dos valores do 3.º quartil e da amplitude interquartílica; ou três vezes inferiores ao valor do 1.º quartil menos a amplitude interquartílica). No exemplo que apresentamos de seguida (Figura 5), não foram observados nenhum destes valores. Partindo de medidas de tendência não-central (quartis), este gráfico permite uma visualização imediata da distribuição da variável.

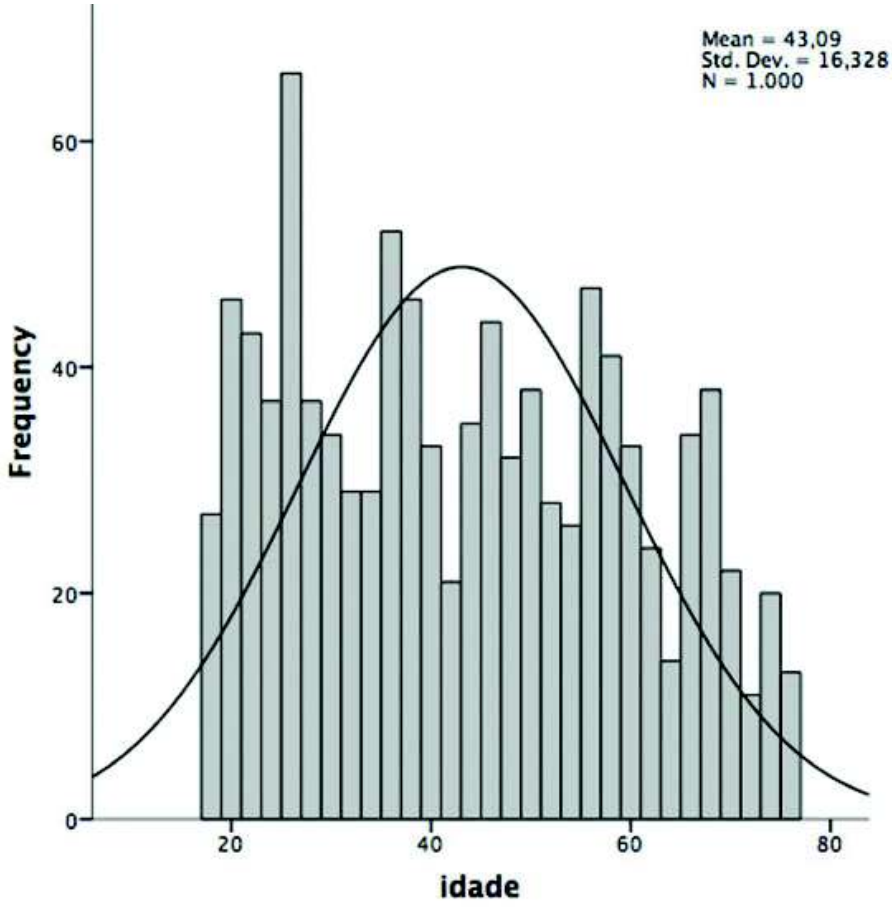
Figura 5. Gráfico de extremos e quartis da variável «idade»



Fonte dos dados: *Violência contra as mulheres*, SociNova (1995).

Por último, destacamos o histograma de frequências, que se constitui como uma representação gráfica das observações organizadas em classes. Quando construído a partir do SPSS, este gráfico permite ainda a apresentação das principais medidas descritivas (como a média e o desvio padrão), e também o ajustamento de uma curva normal à distribuição (Figura 6).

Figura 6. Histograma de frequências da variável «idade»



Fonte dos dados: *Violência contra as mulheres*, SociNova (1995).

Este gráfico permite a visualização da distribuição de variáveis métricas, transmitindo, de forma mais imediata, o seu comportamento, nomeadamente no que diz respeito ao enviesamento.

Análise de dados bivariada

A análise de dados bivariada permite, numa perspectiva descritiva e também inferencial, avaliar a relação entre duas variáveis. Estes procedimentos podem ser importantes por si só (como para perceber a relação entre duas variáveis), ou podem ser úteis enquanto

fase preparatória de análises multivariadas, não sendo estas opções mutuamente exclusivas.

Nas Ciências Sociais, e mais concretamente na Sociologia, os estudos de carácter mais quantitativo recorrem, em grande medida, a dados recolhidos através de inquéritos por questionários e incidem, na grande maioria dos casos, sobre práticas, opiniões, percepções, valores e atitudes. Daqui resulta, pela própria natureza da informação, uma predominância de variáveis de tipo categórico, quer sejam nominais ou ordinais (neste caso, resultantes da aplicação de escalas de tipo Likert). Assim, é menos frequente dispormos de variáveis métricas, pelo que nos concentraremos mais nos exemplos que fazem recurso a variáveis de tipo categórico.

Nos inquéritos à vitimação e aos custos e consequências da violência contra as mulheres, doméstica e de género, por exemplo, é fundamental perceber o comportamento de determinadas variáveis (características sociodemográficas, presença de sintomas ou doenças físicos, estados psicológicos, entre outras), em função do facto de a pessoa ter ou não sido vítima, no sentido de aferir se as realidades observadas são comuns à população em geral ou se decorrem de situações específicas de violência.

Antes de passarmos à exemplificação das metodologias de análise mais frequentemente utilizadas, convém, neste momento, fazer uma distinção importante entre aquilo que é a estatística descritiva e a estatística inferencial. A estatística descritiva, tal como a própria designação indica, remete para uma descrição dos dados de que dispomos, e as conclusões que retiramos aplicam-se somente aos elementos em análise. Tal como define Reis, a estatística descritiva «consiste na recolha, apresentação, análise e interpretação de dados numéricos através da criação de instrumentos adequados: quadros, gráficos e indicadores numéricos» (2009: 15). Já a estatística inferencial permite extrapolar para a população (ou universo) os resultados obtidos através dos dados de uma amostra. Debruçar-nos-emos, em primeiro lugar, sobre as estatísticas bivariadas adequadas a variáveis categóricas, passando de uma abordagem descritiva para uma abordagem inferencial.

Um dos principais, e mais comummente utilizados, instrumentos para o estudo da distribuição conjunta de duas variáveis categóricas são as tabelas de contingência. Isto porque, como fizemos notar anteriormente, a grande maioria das variáveis com que trabalhamos nas Ciências Sociais são desta mesma natureza (nominal ou ordinal). As tabelas de contingência fornecem um conjunto diversificado de informação, que passaremos a analisar, partindo de um exemplo que cruza a vitimação (apenas de violência contra as mulheres) no ano anterior à aplicação do questionário e o estado civil à época (Quadro 9). Consideremos, para já, apenas as frequências absolutas observadas (designadas por *count* no SPSS). Em primeiro lugar, do geral para o particular, conseguimos perceber qual a dimensão da amostra com que estamos a trabalhar, isto é, quantas respostas válidas temos, simultaneamente, às duas variáveis. Neste caso, todos os elementos da amostra inicial têm respostas válidas, tanto no «estado civil», como na «vitimação», correspondendo a 1000 observações (secção do quadro sombreado a preto). De seguida, encontramos informação relativa aos totais marginais da tabela (sombreados a cinzento escuro), e que representam a distribuição de cada uma das variáveis, independentemente da interacção com a outra. Por exemplo, percebemos que, das mil mulheres desta amostra, 255 são solteiras (independentemente de serem ou não vítimas); seguindo a mesma lógica, observamos que 131 das 1000 mulheres inquiridas foram vítimas de violência doméstica no ano anterior à aplicação do questionário, independentemente do seu estado civil. Os totais marginais funcionam, neste sentido, como estatísticas descritivas univariadas de cada uma das variáveis em causa. De facto, nenhuma destas estatísticas acrescenta nada de novo relativamente às tabelas de frequências anteriormente exploradas; a informação efectivamente importante e nova que podemos retirar das tabelas cruzadas são as contagens (valores observados) das combinações das várias categorias que compõem cada uma das variáveis.

Quadro 9. Tabela de contingência das variáveis «vitimação» e «estado civil» (valores observados e percentagens)

v103a estado civil actual * vdom_uano Violência doméstica no último ano Crosstabulation

			vdom_uano Violência doméstica no último ano		Total
			0 não vítima	1 vítima	
v103a estado civil actual	1 solteira	Count	215	40	255
		% within v103a estado civil actual	84,3%	15,7%	100,0%
		% within vdom_uano Violência doméstica no último ano	24,7%	30,5%	25,5%
		% of Total	21,5%	4,0%	25,5%
	2 casada/união de facto	Count	500	76	576
		% within v103a estado civil actual	86,8%	13,2%	100,0%
		% within vdom_uano Violência doméstica no último ano	57,5%	58,0%	57,6%
		% of Total	50,0%	7,6%	57,6%
	3 divorciada/ separada	Count	34	8	42
		% within v103a estado civil actual	81,0%	19,0%	100,0%
		% within vdom_uano Violência doméstica no último ano	3,9%	6,1%	4,2%
		% of Total	3,4%	,8%	4,2%
	4 viúva	Count	120	7	127
		% within v103a estado civil actual	94,5%	5,5%	100,0%
		% within vdom_uano Violência doméstica no último ano	13,8%	5,3%	12,7%
		% of Total	12,0%	,7%	12,7%
Total		Count	869	131	1000
		% within v103a estado civil actual	86,9%	13,1%	100,0%
		% within vdom_uano Violência doméstica no último ano	100,0%	100,0%	100,0%
		% of Total	86,9%	13,1%	100,0%

Fonte dos dados: *Violência contra as mulheres*, SociNova (1995).

Observamos, por exemplo, que, nesta amostra, existem 76 mulheres casadas ou em união de facto que foram vítimas de violência doméstica no ano anterior ao questionário. Mas mais útil do que analisar os valores absolutos, porque não dão conta do peso do valor da célula no conjunto dos dados, ou em grupos específicos, é olhar para os valores percentuais. E aqui poderemos fazer a análise de diferentes formas, recorrendo às percentagens em linha, em coluna ou às percentagens relativamente ao total.

As percentagens em linha representam o peso do valor observado no total da categoria em linha. Por exemplo, e recorrendo aos dados apresentados na tabela cruzada anterior (Quadro 9), o peso relativo das mulheres solteiras vítimas no conjunto das solteiras é de 15,7% (=40/255*100). Assim, observa-se que 15,7% das mulheres solteiras foram vítimas.

Às percentagens em coluna aplica-se o mesmo raciocínio, sendo que, agora, os totais que servem de referência aos cálculos são os totais das colunas. Por exemplo, o peso das mulheres solteiras vítimas no conjunto das vítimas é de 30,5% (=40/131*100), ou seja, 30,5% das vítimas são solteiras.

Por último, podemos ainda analisar as percentagens relativamente ao total da amostra. Estas representam o peso da célula no conjunto dos elementos da amostra (mais especificamente, do total de elementos da amostra que têm respostas válidas em ambas as variáveis). Assim, e ainda recorrendo ao mesmo exemplo, podemos observar que a amostra é composta por 4,0% (=40/1000*100) de mulheres que são solteiras e vítimas de violência doméstica.

A escolha de qual dos tipos de percentagem a analisar depende, exclusivamente, do investigador, e decorre dos objectivos do seu trabalho e do seu objecto de estudo. No caso que aqui apresentamos, o nosso interesse é o de conhecer as características das mulheres que foram vítimas de violência doméstica, pelo que dedicaríamos a nossa atenção à análise das percentagens em coluna. Esta informação diz-nos como se distribuem as mulheres vítimas (e as não vítimas) por cada um dos estados civis, permitindo, de uma forma ainda muito primária, identificar os grupos mais vulneráveis a este fenómeno.

Mas devemos acrescentar alguma informação adicional a esta leitura. Olhando para as percentagens em coluna (Quadro 9), identificaríamos as casadas como o grupo onde encontramos mais vítimas de violência (58,0%). Contudo, significará este valor que são as casadas o grupo onde este fenómeno mais ocorre? Poderíamos afirmar que sim, se os grupos de mulheres em cada estado civil fosse de semelhante dimensão – o que não se verifica. É verdade que 58% das vítimas são casadas, mas também é verdade que, na nossa amostra, mais de metade das inquiridas são casadas (57,6%). Será que é a dimensão do grupo na amostra que faz inflacionar o seu peso no conjunto das vítimas? A melhor forma de fazer esta verificação é comparar as percentagens, neste caso, em coluna, com a respectiva percentagem marginal, o que nos dá a noção do peso do cruzamento relativamente à dimensão do grupo. Sabendo, então, que as mulheres casadas representam 57,6% da nossa amostra, podemos ver que o diferencial entre vítimas e não vítimas não é particularmente expressivo, ou seja, o peso das casadas nas não vítimas (57,5%) está muito ligeiramente abaixo da proporção de mulheres casadas como um todo (57,6%), e o peso das vítimas casadas (58%) está ligeiramente acima do conjunto das casadas na amostra. Feita então a leitura nestes moldes, percebe-se que a proporção de vítimas casadas e de não vítimas casadas não difere grandemente, ainda que a significância estatística destas relações (ou da ausência delas) se faça a partir de uma perspectiva inferencial, nomeadamente através do teste de independência do Qui-Quadrado e da avaliação dos resíduos, processo que descreveremos de seguida.

O teste de independência de Qui-Quadrado (χ^2) vai permitir perceber se existe, ou não, uma relação significativa entre a vitimação e o estado civil na população, ou seja, «se a frequência com que os elementos da amostra se repartem pelas classes de uma variável nominal categorizada é ou não aleatória» (Marôco, 2014: 113). Para os nossos dados, os resultados são os apresentados de seguida[2]:

2 Para uma explicação mais aprofundada do procedimento estatístico, ver e.g., Marôco (2014); para o procedimento em SPSS, ver e.g., Laureano (2013).

Quadro 10. Resultados do teste de para as variáveis «vitimação» e «estado civil»

Chi-Square Tests			
	Value	**df**	**Asymp. Sig. (2-sided)**
Pearson Chi-Square	9,232[a]	3	,026
Likelihood Ratio	10,567	3	,014
Linear-by-Linear Association	5,967	1	,015
N of Valid Cases	1000		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 5,50.

Fonte dos dados: *Violência contra as mulheres*, SociNova (1995).

Definidas as hipóteses em teste, hipótese nula (H_0) e hipótese alternativa (H_a) (H_0: a vitimação e o estado civil são independentes na população; H_a: a vitimação e o estado civil não são independentes na população), podemos concluir pela rejeição da hipótese nula (já que *sig.* = 0,026 ≤ α = 0,05, para um nível de confiança de 95%). Podemos então dizer que foi verificada a existência de uma relação significativa entre o estado civil das mulheres e terem ou não sido vítimas de violência doméstica ($\chi^2_{(3)} = 9{,}232$, $p = 0{,}026$).

Complementarmente à análise da independência, no caso de esta ser rejeitada, podemos ainda aferir a intensidade dessa mesma relação. Apesar de o valor da estatística de teste do Qui² nos dar alguma noção sobre isto (quanto maior o Qui², maior o afastamento da hipótese de independência), não é adequado utilizá-la para este efeito, uma vez que não varia num intervalo fixo, dependendo da dimensão da amostra (Reis, Melo, Andrade, & Calapez, 2016). Em alternativa, devem ser usadas medidas que, baseadas no Qui², forneçam essa mesma informação de uma forma padronizada. Exemplo disto é o V de Cramer[3], medida de associação que mede a intensidade da relação entre as variáveis. Assim sendo, só faz sentido ser aplicado após a rejeição da hipótese nula no teste do Qui² – portanto, quando se conclui pela rejeição da independência. O V de Cramer varia, então, entre 0 e 1, correspondendo 0 a uma relação nula e 1 a uma relação perfeita. No caso que

3 A medida do V de Cramer é dada por $V=\sqrt{\frac{x^2/n}{q\text{-}1}}$, onde *q* corresponde ao menor valor de linhas e colunas da tabela cruzada.

estamos a analisar, tendo concluído pela existência de uma relação significativa entre a vitimação e o estado civil, vamos aferir a sua intensidade. Como se observa (Quadro 11), a relação entre as duas variáveis, ainda que exista, é muito fraca, já que o coeficiente está muito próximo de 0 (V = 0,096).

Quadro 11. Resultado da medida de associação V de Cramer para a relação estabelecida entre «vitimação» e «estado civil»

Symmetric Measures			
		Value	**Approx. Sig.**
Nominal by Nominal	Phi	,096	,026
	Cramer's V	,096	,026
N of Valid Cases		1000	

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

Fonte dos dados: *Violência contra as mulheres*, SociNova (1995).

Não sabemos, ainda, nada acerca da configuração desta relação. Para tal, deveremos recorrer à análise dos resíduos estandardizados e ajustados[4], de modo a identificar quais as células que contribuem para a significância da relação global entre as variáveis[5] (Agresti, 2013). Como se observa nos resultados apresentados de seguida (Quadro 12), encontramos uma relação significativa apenas entre

4 Os resíduos representam a diferença entre o valor observado (VO) e o valor esperado (VE) em caso de independência ($VO - VE$). Uma vez que valores esperados elevados vão produzir resíduos elevados (a magnitude dos resíduos não depende apenas da força da associação, mas também da dimensão das categorias), eles devem ser estandardizados [$(VO - VE)/\sqrt{VE}$]. O resíduo estandardizado ajustado (mais comummente utilizado por ser ajustado pelo erro padrão, porque nem sempre os resíduos estandardizados se ajustam convenientemente à distribuição normal padrão) é dado por $\frac{(VO - VE)}{\sqrt{VE \cdot (1 - TML/n) \cdot (1 - TMC/n)}}$, onde TML corresponde ao total marginal da linha e TMC ao total marginal da coluna (Agresti, 2013; Sharpe, 2015).

5 Tendo em conta que os resíduos ajustados seguem uma distribuição normal padrão, e sabendo que o valor crítico de z corresponde a 1,96 valores de resíduo ajustado superiores a 1,96 (ou inferiores a -1,96) são indicativos de uma falta de ajustamento à H_0, pelo que podemos considerá-los significativos e, por extensão, também a relação estabelecida entre as duas categorias.

ser-se viúva e não se ser vítima (*resíduo* = 2,7)[6], o que sugere que é sobretudo esta categoria que torna significativa a relação entre a vitimação e o estado civil (aferida através dos resultados do teste de χ^2). Assim, podemos concluir que, globalmente, as vítimas não apresentam um padrão particular relativamente ao estado civil.

Quadro 12. Tabela de contingência das variáveis «vitimação» e «estado civil» (valores observados, valores esperados e resíduos ajustados)

v103a estado civil actual * vdom_uano Violência doméstica no último ano Crosstabulation

			vdom_uano Violência doméstica no último ano		Total
			0 não vítima	1 vítima	
v103a estado civil actual	1 solteira	Count	215	40	255
		Expected Count	221,6	33,4	255,0
		Adjusted Residual	-1,4	1,4	
	2 casada/união de facto	Count	500	76	576
		Expected Count	500,5	75,5	576,0
		Adjusted Residual	-,1	,1	
	3 divorciada/separada	Count	34	8	42
		Expected Count	36,5	5,5	42,0
		Adjusted Residual	-1,2	1,2	
	4 viúva	Count	120	7	127
		Expected Count	110,4	16,6	127,0
		Adjusted Residual	2,7	-2,7	
Total		Count	869	131	1000
		Expected Count	869,0	131,0	1000,0

Fonte dos dados: *Violência contra as mulheres*, SociNova (1995).

6 Verifica-se, igualmente, uma relação significativa entre ser-se viúva e ser-se vítima mas, sendo o resíduo ajustado negativo, significa que há uma forte probabilidade de as categorias não ocorrerem em conjunto. Como se trata de uma conclusão redundante, não é hábito proceder-se à análise dos resíduos negativos.

Ainda uma nota final a propósito do teste de Qui-Quadrado. Para poder confiar nos resultados obtidos, é necessário que algumas condições de aplicação estejam garantidas. Assim, antes da análise dos resultados, há que verificar os pressupostos de aplicação do teste, que são: não existir mais de 20% de células com valor esperado inferior a 5; e não existir nenhuma célula com valor esperado inferior a 1. Estes pressupostos podem ser verificados «manualmente», através de uma tabela cruzada a que adicionamos os valores esperados (Quadro 12). Porém, e sobretudo se estivermos a analisar tabelas muito grandes, este método revela-se pouco prático. O SPSS fornece, aquando da apresentação do *output* do Qui-Quadrado, a informação necessária para este procedimento. Assim, e atentando à nota de rodapé apresentada para a estatística de teste (Quadro 10), podemos observar que nenhuma célula tem valor esperado inferior a 5[7] (note-se que seria aceitável ter até 20% das células com estes valores) e nenhuma célula tem um valor esperado inferior a 1 (*The minimum expected count is 5,50*). Estando assim cumpridos ambos os pressupostos, poderíamos avançar com a análise dos resultados do teste. Caso esta situação não se verificasse, ou seja, caso não cumpríssemos ambos os pressupostos, deveríamos recorrer ao teste exacto de Fisher. Este teste, inicialmente desenvolvido para tabelas 2x2, deve ser utilizado quando não estão cumpridos os pressupostos do teste de Qui-Quadrado. Contudo, o SPSS só o disponibiliza, no módulo *standard* SPSS Statistics, para tabelas 2x2; para tabelas com outras dimensões, é necessário ter o módulo *Exact Tests* instalado (Marôco, 2014).

Vejamos agora alguns exemplos das estatísticas descritivas (para variáveis métricas) anteriormente apresentadas, mas agora sob uma perspectiva bivariada (Quadro 13).

7 *«0 cells (0,0%) have expected count less than 5.»*

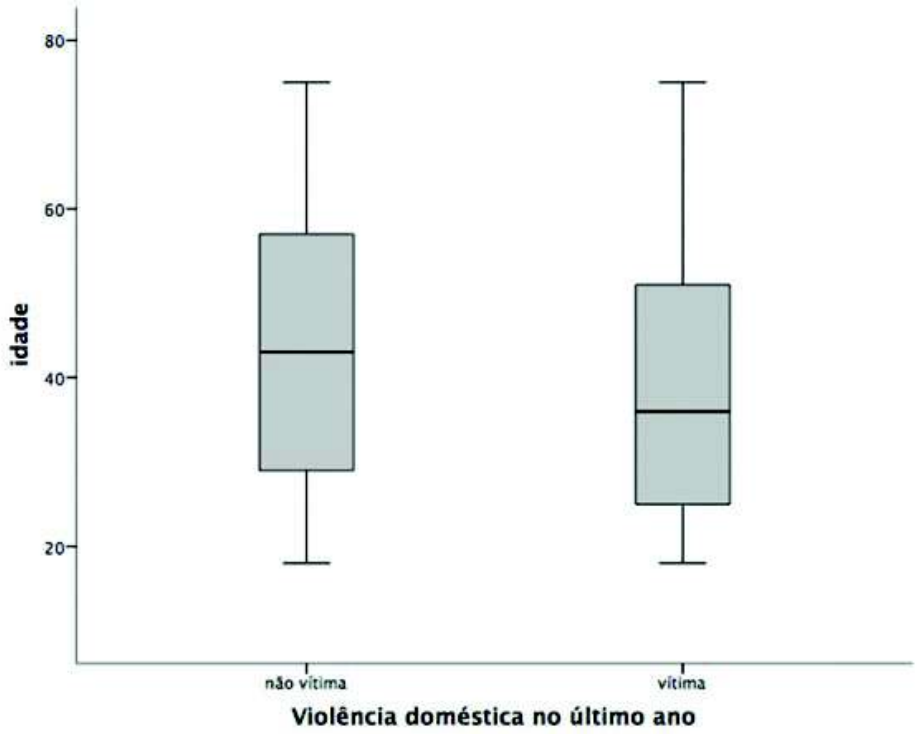
Quadro 13. Estatísticas descritivas da variável «idade» por «vitimação»

Descriptives				
vdom_uano Violência doméstica no último ano			Statistic	Std. Error
v102 idade0 não vítima	Mean		43,70	,557
	95% Confidence	Lower Bound	42,61	
		Upper Bound	44,80	
	5% Trimmed Mean		43,46	
	Median		43,00	
	Variance		269,283	
	Std. Deviation		16,410	
	Minimum		18	
	Maximum		75	
	Range		57	
	Interquartile Range		29	
	Skewness		,155	,083
	Kurtosis		-1,151	,166
1 vítima	Mean		39,05	1,330
	95% Confidence	Lower Bound	36,42	
		Upper Bound	41,69	
	5% Trimmed Mean		38,39	
	Median		36,00	
	Variance		231,882	
	Std. Deviation		15,228	
	Minimum		18	
	Maximum		75	
	Range		57	
	Interquartile Range		26	
	Skewness		,411	,212
	Kurtosis		-,897	,420

Fonte dos dados: *Violência contra as mulheres*, SociNova (1995).

Observam-se, portanto, as várias estatísticas descritivas disponíveis para cada um dos grupos definidos. Por exemplo, a média das idades das mulheres que foram vítimas de violência doméstica no ano anterior à aplicação do questionário é de 39,05 anos, e das que não foram vítimas é de 43,70 anos. Podemos ainda explorar graficamente se vítimas e não vítimas apresentam uma distribuição etária diferenciada, através do gráfico de extremos e quartis, já apresentado anteriormente. Observa-se então (Figura 7) que as mulheres vítimas parecem ser mais jovens do que as não vítimas, já que apresentam uma mediana e um 3.º quartil inferiores aos das não vítimas. Estando ainda numa fase de exploração dos dados relativamente a este indicador, não devemos olhar para estes resultados na perspectiva da extrapolação para a população.

Figura 7. Gráfico de extremos e quartis da variável «idade» por grupos (vítimas/não vítimas)



Fonte dos dados: *Violência contra as mulheres*, SociNova (1995).

Esta leitura, também apoiada nas estatísticas descritivas atrás apresentadas, justifica a aplicação de um teste que possa avaliar se as médias das idades destes dois grupos são ou não significativamente diferentes na população. Para tal, devemos proceder à aplicação do teste *t* à comparação de médias para duas amostras independentes[8], utilizável «quando se tem uma variável quantitativa (dependente) e se pretende comparar a sua média em dois grupos populacionais independentes definidos por uma variável qualitativa (independente)» (Laureano, 2013: 28). Contudo, é necessário, para a sua aplicação, que sejam cumpridos dois pressupostos: (1) independência dos grupos, onde é necessário garantir que os grupos em causa são mutuamente exclusivos; e (2) os grupos serem retirados de uma população com distribuição normal, pelo que é necessário garantir que a variável dependente segue uma distribuição normal na população para cada um dos grupos[9]. Retomando o exemplo anterior, onde pretendemos perceber se as médias das idades de vítimas e não vítimas são significativamente diferentes, vamos, em primeiro lugar, abordar a questão dos pressupostos. Se a verificação do primeiro pressuposto decorre da forma como a variável foi construída na base de dados, e isso pode ser avaliado facilmente (e, neste caso, está verificado, uma vez que os grupos definidos pela variável «vitimação» são mutuamente exclusivos), para a validação do segundo, é necessário recorrer ao teste de Kolmogorov-Smirnov ou de Shapiro-Wilk, já anteriormente abordados (Quadro 8). Assim, deve aplicar-se este procedimento para cada um dos grupos, pelo que é necessária a formulação de dois conjuntos de hipóteses distintos: para o grupo das vítimas, com H_0: a idade das vítimas segue uma distribuição normal na população e H_a: a idade das vítimas não segue uma distribuição normal na população; e para o grupo das não vítimas, onde H_0: a idade das não vítimas segue uma distribuição normal

8 No caso de trabalharmos com amostras emparelhadas e querermos testar esta hipótese, deve usar-se o teste *t* para amostras emparelhadas.

9 O teste *t* é robusto à violação do pressuposto da normalidade da distribuição no caso de amostras de grande dimensão (decorrente do teorema do limite central) e de não serem particularmente enviesadas ou achatadas (Field, 2013; Marôco, 2014). Nesse caso, deve recorrer-se ao teste de Mann-Whitney, alternativa não-paramétrica ao teste *t*.

na população e H_a: a idade das não vítimas não segue uma distribuição normal na população.

Quadro 14. Testes à normalidade da distribuição da variável «idade» por grupos (vítimas/não vítimas)

Tests of Normality

vdom_uano Violência doméstica no último ano		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
v102 idade	0 não vítima	,073	869	,000	,954	869	,000
	1 vítima	,136	131	,000	,934	131	,000

a. Lilliefors Significance Correction

Fonte dos dados: *Violência contra as mulheres*, SociNova (1995).

Tendo ambos os grupos dimensões superiores a 50 elementos, devemos recorrer aos resultados do teste de Kolmogorov-Smirnov, que nos aponta (Quadro 14) para a rejeição de ambas as hipóteses nulas ($KS_{(131)} = 0{,}136$, $p < 0{,}001$ e $KS_{(869)} = 0{,}073$, $p < 0{,}001$, para vítimas e não vítimas, respectivamente), pelo que concluímos que os grupos não provêm de uma população com distribuição normal, violando, assim, este pressuposto de aplicação.

À partida, não deveríamos aplicar o teste *t*, já que não estão cumpridos os pressupostos, mas, como vimos anteriormente, e invocando o teorema do limite central, podemos assumir que este não se constituirá como um problema na estimação dos parâmetros em estudo. Em todo o caso, daremos mais adiante um exemplo da aplicação da alternativa não-paramétrica a este mesmo problema.

Aplicado, então, o teste *t* à comparação de médias de amostras independentes, obtemos os seguintes resultados. Como se observa (Quadro 15), são gerados dois testes *t*: um calculado assumindo que as variâncias são iguais nos dois grupos (*equal variances assumed*) e outro assumindo que as variâncias não são iguais nos dois grupos (*equal variances not assumed*). A decisão sobre qual dos testes *t* interpretar deve basear-se no teste de Levene, teste à homo-

geneidade de variâncias (H_0: a variância da idade das vítimas é igual à variância da idade das não vitimas; H_a: a variância da idade das vítimas não é igual à variância da idade das não vitimas). Face aos resultados obtidos (Quadro 15), decidimos pela não rejeição da hipótese nula em teste (F = 1,243, p = 0,265), podendo afirmar, então, que as variâncias são iguais nos dois grupos, pelo que interpretaremos os resultados do teste t que foi calculado com a assunção de igualdade de variâncias.

Quadro 15. Resultados do teste *t* à comparação das médias das idades de vítimas e não vítimas

Independent Samples Test										
		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
v102 idade	Equal variances assumed	1,243	,265	3,051	998	,002	4,650	1,524	1,659	7,640
	Equal variances not assumed			3,224	178,680	,002	4,650	1,442	1,804	7,496

Fonte dos dados: *Violência contra as mulheres*, SociNova (1995).

Observa-se então (Quadro 15) que existem evidências estatísticas para afirmar que vítimas e não vítimas têm idades (médias) significativamente diferentes, já que se rejeita a hipótese nula em teste[10] ($t_{(998)}$ = 3,051, p = 0,002). Verifica-se ainda, de acordo com

10 Hipóteses em teste no teste t: H_0: a média das idades das não vítimas é igual à média das idades das vítimas; H_a: a média das idades das não vítimas é diferente da média das idades das vítimas.

o intervalo de confiança de 95% para a diferença das médias das idades de]1,659; 7,640[, que as mulheres não vítimas são, em média, entre 1,7 e 7,7 anos mais velhas do que as vítimas[11]. Sabendo que os métodos estatísticos são aqui usados para responder a questões levantadas durante o processo de investigação, devemos pensar neste resultado como um ponto de partida para uma exploração mais profunda do tema, tanto mais que, nos inquéritos nacionais realizados posteriormente, há uma transversalidade da prevalência da vitimação em relação à idade. Se as mulheres vítimas são, em média, mais novas do que as não vítimas, as perguntas que se seguem devem remeter para a procura de uma explicação deste facto. Será que as mulheres mais novas estão mais expostas aos actos de violência e/ou será que têm um menor nível de aceitação destas práticas, pelo que falam delas mais abertamente? Ou ainda, sendo este o primeiro inquérito nacional à violência contra as mulheres, aplicado num momento de muito menor visibilidade social do fenómeno (por relação à actualidade), pode admitir-se a hipótese de um maior silenciamento das pessoas mais velhas.

Caso preferíssemos utilizar a alternativa não-paramétrica ao teste t, recorreríamos ao teste de Mann-Whitney, que, para além de ser usado em alternativa ao teste *t* quando não estão cumpridos os pressupostos e não se quer evocar o teorema do limite central, é «adequado para comparar as funções de distribuição de uma variável pelo menos ordinal medida em duas amostras independentes» (Marôco, 2014: 321). Uma vez que é possível a aplicação deste teste a variáveis de tipo ordinal, não são testadas as médias da variável em estudo, mas sim as médias das ordenações (*mean rank*) da variável dependente (ver, e.g., Reis *et al.*, 2016). As hipóteses em teste são semelhantes às consideradas no teste *t*, sendo agora H_0: a média das ordenações das idades das não vítimas é igual à média das ordenações das idades das vítimas; e H_a: a média das ordenações das idades das não vítimas é diferente da média das ordenações das idades das vítimas. À semelhança dos

11 Esta interpretação decorre das categorias que definimos como grupo 1 e grupo 2. Nesta análise, considerámos as não vítimas como grupo 1 e as vítimas como grupo 2. Considerando que a diferença das médias é dada por $grupo_1 - grupo_2$, uma diferença positiva significa que o grupo 1 apresenta valores mais altos que o grupo 2.

testes apresentados até agora, opta-se pela rejeição da hipótese nula quando *sig.* < 0,05. Como se observa (Quadro 16), as não vítimas apresentam uma média das ordenações (MRk_0 = 511,15) mais elevada do que as vítimas (MRk_1 = 429,87), e essa diferença é significativa na população (U = 47667,50, p = 0,003) (Quadro 17). Assim sendo, a conclusão é semelhante à retirada através da aplicação do teste *t*, ou seja, de que as mulheres vítimas são significativamente mais novas que as mulheres não vítimas.

Quadro 16. *Ranks* da variável «idade» para os grupos «não vítima» e «vítima»

Ranks				
vdom_uano Violência doméstica no último ano		**N**	**Mean Rank**	**Sum of Ranks**
v102 idade	0 não vítima	869	511,15	444186,50
	1 vítima	131	429,87	56313,50
	Total	**1000**		

Fonte dos dados: *Violência contra as mulheres*, SociNova (1995).

Quadro 17. Resultados do teste de Mann-Whitney à comparação das ordenações das idades das vítimas e das não vítimas

Test Statistics[a]		
v102 idade	Mann-Whitney U	47667,500
	Wilcoxon W	56313,500
	Z	-3,003
	Asymp. Sig. (2-tailed)	,003

a. Grouping Variable: vdom_uano Violência doméstica no último ano

Fonte dos dados: *Violência contra as mulheres*, SociNova (1995).

Um outro teste importante, no âmbito das análises estatísticas bivariadas, é o cálculo do coeficiente de correlação. Como vimos anteriormente, o teste do Qui^2 destina-se a analisar a relação entre variáveis de tipo nominal (ou tratadas como nominais).

Mas como aferir a relação que se estabelece entre variáveis de outras naturezas? Por exemplo, se quisermos perceber se existe uma relação entre o número de actos de violência psicológica e o de violência física sofridos (ambas variáveis métricas)? E se quisermos perceber se existe alguma relação entre o número de actos de violência psicológica reportados e o rendimento do agregado familiar (medido em escalões de rendimento, fazendo com que tenhamos uma variável métrica e outra ordinal)? Como estas variáveis foram medidas com escalas que contêm mais informação e, no mínimo, permitem a ordenação entre categorias (ordinais) e, no máximo, a medida exacta da distância entre elas (métricas), é aconselhável que se usem testes estatísticos mais robustos, que permitem ir mais longe na análise da natureza e intensidade da relação entre variáveis, pelo que não faz sentido aplicar um teste de Qui^2. Para este tipo de situações, temos ao nosso dispor outros instrumentos mais adequados, como são os coeficientes de correlação, que medem o grau de associação linear entre duas variáveis. Chamamos a atenção para o facto de, sendo este um coeficiente de correlação linear, apenas detecta relações deste tipo. Os coeficientes de correlação mais comummente utilizados são o R de Pearson e o ρ de Spearman. O coeficiente de correlação de Pearson[12] é adequado para a correlação entre duas variáveis métricas, ao passo que o coeficiente de correlação de Spearman[13] deve ser utilizado para variáveis ordinais (ou quando uma delas é ordinal e a outra métrica). Apesar de terem formas de cálculo bastante diferentes (o primeiro baseia-se na correlação entre os valores das variáveis e o segundo na das ordenações das observações), a sua análise é em tudo semelhante. Ambos variam entre -1 e 1, correspondendo |1| a correlações perfeitas e 0 a correlações nulas. A análise dos coeficientes de correlação deve fazer-se através de dois indicadores: a intensidade e o sinal da relação entre as variáveis. Valores superiores, em módulo, a 0,5

12 O coeficiente de correlação linear de Pearson é dado por $r=\frac{n\cdot\sum X_i\cdot Y_i-\sum X_i\cdot\sum Y_i}{\sqrt{\left[n\cdot\sum X_i-\left(\sum X_i\right)\right]\cdot\left[n\cdot\sum Y_i-\left(\sum Y_i\right)\right]}}$

13 O coeficiente de correlação linear de Spearman é dado por $\rho=1-\frac{6\cdot\sum d}{n\cdot(n-1)}$, onde *d* corresponde à diferença, para cada observação, entre o número de ordem atribuído na variável *X* e o atribuído na variável *Y*.

representam uma intensidade elevada da correlação entre as duas variáveis; quando são inferiores a 0,5, em módulo, a correlação é baixa, ainda que se devam ter em consideração os indicadores do SPSS, que podem validar valores menos expressivos. O sinal da correlação, positivo ou negativo, indica se as variáveis variam no mesmo sentido ou em sentido contrário. Se o sinal é positivo (+), isso significa que variam no mesmo sentido, ou seja, se os valores de uma variável aumentam, os da outra também, ou então que, quando diminuem numa, também diminuem na outra. Se o sinal é negativo (-), variam em sentido contrário, ou seja, se uma variável aumenta, a outra diminui e vice-versa. Sendo também estes testes de hipóteses, há que, em primeiro lugar, formular as hipóteses em teste. Assim sendo, e em ambos os casos (R e ρ), H_0: a correlação é nula na população; H_a: a correlação não é nula na população.

Vejamos dois exemplos. Sob a hipótese teórica de que a violência contra as mulheres tem um pano de fundo estrutural e da coexistência de vários tipos de violência, pretendemos perceber se o número de actos de violência física e o número de actos de violência psicológica sofridos estão ou não relacionados. Para tal, e sendo que estamos perante duas variáveis métricas, recorremos ao coeficiente de correlação linear de Pearson. Os resultados obtidos (Quadro 18) permitem-nos, por um lado, perceber que a relação que se estabelece entre as duas variáveis é significativa e, por outro lado, que a relação é forte e positiva (r = 0,662, p < 0,001). Assim, podemos afirmar que as mulheres que foram vítimas de muitos actos de violência física também o foram de violência psicológica.

Quadro 18. Resultados dos testes de correlação de Pearson entre o número de actos de violência física e o número de actos de violência psicológica sofridos

Correlations			
		ac_pfis Número de actos de violência física	**ac_ppsi Número de actos de violência psicológica**
ac_pfis Número de actos de violência física	Pearson Correlation	1	,662**
	Sig. (2-tailed)		,000
	N	1000	1000
ac_ppsi Número de actos de violência psicológica	Pearson Correlation	,662**	1
	Sig. (2-tailed)	,000	
	N	1000	1000

**. Correlation is significant at the 0.01 level (2-tailed).

Fonte dos dados: *Violência contra as mulheres*, SociNova (1995).

No sentido de perceber se existe alguma relação entre os níveis de rendimento do agregado familiar e o número de actos de violência psicológica sofridos pelas mulheres, procedemos à aplicação do coeficiente de correlação de Spearman (tendo em conta que estamos perante uma variável ordinal e uma métrica). Perante os resultados (Quadro 19), somos levados a rejeitar a hipótese nula em teste (H_0: a correlação entre o rendimento do agregado familiar e o número de actos de violência psicológica sofridos é nula na população; H_a: a correlação entre o rendimento do agregado familiar e o número de actos de violência psicológica sofridos não é nula na população), e ainda a afirmar que a relação que se estabelece é muito fraca e negativa ($\rho = -0{,}139$, $p < 0{,}001$). Poderíamos então dizer que, quanto menor fosse o rendimento do agregado, de mais actos de violência psicológica as mulheres teriam sido vítimas. Contudo, e considerando que esta relação é bastante fraca, não devemos retirar as conclusões nestes termos, mas sim explorar esta dimensão através de métodos multivariados. Aliás, como observaram Lourenço, Lisboa e Pais (1997), através de uma análise factorial, a violência psicológica não parece estar particularmente associada a nenhum estrato social.

Quadro 19. Resultados dos testes de correlação de Spearman entre o número de actos de violência psicológica sofrida e os escalões de rendimento do agregado familiar

Correlations				
			ac_ppsi Número de actos de violência psicológica	**v112 escalão de rendimentos do agregado familiar**
Spearman's rho	ac_ppsi Número de actos de violência psicológica	Correlation Coefficient	1,000	-,139**
		Sig. (2-tailed)		,000
		N	1000	962
	v112 escalão de rendimentos do agregado familiar	Correlation Coefficient	-,139**	1,000
		Sig. (2-tailed)	,000	
		N	962	962

**. Correlation is significant at the 0.01 level (2-tailed).

Fonte dos dados: *Violência contra as mulheres*, SociNova (1995).

Análise de dados multivariada

Tal como descrevem Hair *et al.*, «Multivariate analysis refers to all statistical techniques that simultaneously analyze multiple measurements on individuals or objects under investigation. Thus, any simultaneous analysis of more than two variables can be loosely considered multivariate analysis» (Hair, Hair, Black, Babin, & Anderson, 2013: 4). Porque os fenómenos, nomeadamente os sociais, não podem ser explicados apenas pela relação entre duas grandezas, a utilização de técnicas que permitam abarcar um maior número de factores explicativos torna-se imperativa na maioria dos casos. De facto, a sua popularidade deve-se sobretudo à crescente complexificação da investigação actual (Tabachnick & Fidell, 2013). Tal como qualquer outro tipo de análise, as técnicas de análise multivariada a mobilizar dependem do tipo de informação que pretendemos retirar (portanto, do problema ao qual queremos dar resposta) e do tipo de variáveis que temos ao nosso dispor.

Apresentaremos aqui alguns exemplos de uso mais comum na área das Ciências Sociais. Não sendo de todo exaustivos rela-

tivamente à multiplicidade de técnicas disponíveis, temos como principal objectivo abarcar técnicas com diferentes finalidades e que recorrem a diferentes tipos de variáveis, sobretudo numa perspectiva de variedade.

Análise em Componentes Principais (ACP)

A Análise em Componentes Principais (ACP) constitui-se como uma técnica exploratória de análise multivariada de dados quantitativos, expressos por variáveis métricas, «que transforma um conjunto de variáveis correlacionadas num conjunto menor de variáveis independentes, combinações lineares das variáveis originais, designadas por "componentes principais"» (Marôco, 2014: 455). Tendo como objectivo reduzir a complexidade dos dados, permite a «compreensão dos processos de comportamento dos indivíduos, através da identificação e interpretação dos factores subjacentes» (Reis, 2001: 255).

Assim, a ACP permite que, partindo de *m* variáveis de *input* com algum grau de multicolinearidade (i.e., correlacionadas entre si), se definam *p* novas variáveis (com $p < m$) não correlacionadas ou ortogonais, que se denominam componentes principais. A ACP não explica as correlações entre as variáveis, mas antes encontra combinações lineares entre as variáveis iniciais, que expliquem o máximo possível da variação dos dados (Tabachnick & Fidell, 2013). Considerem-se três variáveis de *input* X_1, X_2 e X_3. A partir destas, é possível criar três componentes principais (cp_1, cp_2 e cp_3), que correspondem a combinações lineares das três variáveis originais, da seguinte forma:

$$\begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \end{array} \begin{array}{c} \begin{array}{ccc} cp_1 & cp_2 & cp_3 \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \end{array}$$

tal que $cp_1=a_{11}\cdot X_1+a_{21}\cdot X_2+a_{31}\cdot X_3$; $cp_2=a_{12}\cdot X_1+a_{22}\cdot X_2+a_{32}\cdot X_3$; $cp_3=a_{13}\cdot X_1+a_{23}\cdot X_2+a_{33}\cdot X_3$.

À primeira componente extraída está sempre associada a maior proporção de variância das variáveis originais, à segunda componente a maior proporção da variância que ainda sobra, e assim sucessivamente, até que, cumulativamente, com a última componente (que tem associada a menor parcela de variância), toda a variância fica explicada. Mas ao aplicar-se uma ACP com o objectivo de identificar dimensões subjacentes às variáveis de *input* (dimensões ou conceitos latentes) ou de reduzir a multidimensionalidade formada pelas variáveis originais, está em causa determinar um número de componentes inferior ao número de variáveis de *input*. Porque reduzir a multidimensionalidade dos dados implica perder informação, importa saber se aquilo que se ganha em termos de interpretabilidade compensa o que se perde em informação (relação custo/benefício). E essa avaliação é feita através de indicadores estatísticos e da interpretação dos agrupamentos de variáveis sugeridos: importa, em primeiro lugar, que as componentes tenham sentido interpretativo e, depois, que sejam estatisticamente válidas (Tabachnick & Fidell, 2013).

Para que a ACP possa ser aplicada, as variáveis de partida deverão ser de tipo quantitativo, ainda que seja comum a utilização de variáveis ordinais com escalas a partir dos cinco pontos (Carifio & Perla, 2008; Ho, 2006).

As desigualdades de género não estão patentes apenas na dimensão da violência. Elas colocam-se, entre outras, também ao nível da igualdade de oportunidades no acesso a cargos de decisão política, que é, igualmente, um indicador da qualidade da democracia. Nesse sentido, e porque as percepções e as atitudes face ao funcionamento do sistema político podem constituir-se como uma condicionante à participação nas elites políticas, seria importante explorar se homens e mulheres revelam entendimentos e posicionamentos diferenciados a este respeito. O *European Social Survey* (ESS)[14] contempla, entre tantas outras dimensões, um conjunto de questões relativas ao posicionamento face a valores e práticas da cidadania, ao funcionamento do

14 O *European Social Survey* é um inquérito conduzido a cada dois anos, com o objectivo de medir atitudes, valores e comportamentos da população de mais de 30 países europeus. A amostra é representativa, a nível nacional, das pessoas maiores de 15 anos, residentes em agregados familiares privados. Em Portugal, a implementação deste inquérito é da responsabilidade do consórcio ICS-UL/ISCTE-IUL.

sistema eleitoral, à relação com instituições governamentais e organizações partidárias e ao funcionamento dos canais de comunicação na sociedade, no âmbito da política. Sendo que o nosso objectivo é comparar as atitudes e as percepções de homens e de mulheres, e considerando que são 12 as variáveis de interesse, podemos pensar em agregar a informação criando indicadores mais genéricos e fazer as análises pretendidas a partir daí. Recorremos então aos dados da população portuguesa do ESS (6.ª ronda, edição 2.0, 2012), relativos ao ano de 2012[15], para realizar uma ACP, com a qual pretendemos transformar as variáveis de partida (Quadro 20) num conjunto mais reduzido de componentes.

Quadro 20. Variáveis de partida para a realização da ACP

Variável	Escala de medida
Em Portugal, os direitos das minorias são protegidos	0 = «não se aplica nada» a 10 = «aplica-se totalmente»
Em Portugal, os cidadãos têm a última palavra nos assuntos políticos mais importantes votando diretamente sobre eles em referendos	
Em Portugal, os partidos do governo que fazem um mau trabalho são castigados nas eleições	
Em Portugal, o governo protege todos os cidadãos da pobreza	
Em Portugal, o governo explica as suas decisões aos eleitores	
Em Portugal, o governo toma medidas para reduzir as diferenças nos níveis de rendimento	
Em Portugal, os políticos têm em conta as opiniões de outros governos Europeus antes de tomarem decisões	
Em Portugal, as eleições legislativas são livres e justas	
Em Portugal, os diferentes partidos políticos apresentam alternativas claras entre si	
Em Portugal, os partidos da oposição são livres para criticar o governo	
Em Portugal, a comunicação social é livre para criticar o governo	
Em Portugal, a comunicação social dá aos cidadãos informação correcta para avaliar o governo	

15 A amostra é constituída por 1522 pessoas, que responderam de forma válida a todas as 12 questões consideradas.

Em primeiro lugar, é necessário avaliar a adequabilidade do procedimento aos dados. Considerando que o objectivo da ACP é o de agrupar variáveis com base no que elas têm de redundante, tem de existir algum grau de correlação entre as variáveis de partida. E isto é avaliado através da medida de Kaiser-Meyer-Olkin (KMO), que quantifica o nível de intercorrelações entre as variáveis e é dado por $KMO=\frac{\sum_i\sum_j r_{ij}^2}{\sum_i\sum_j r_{ij}^2+\sum_i\sum_j a_{ij}^2}$, onde: r_{ij} corresponde ao coeficiente de correlação observado entre as variáveis *i* e *j*; e a_{ij} corresponde ao coeficiente de correlação parcial entre as variáveis *i* e *j,* que é também «uma estimativa das correlações entre os factores» (Reis, 2001: 279). Uma vez que as componentes são independentes, espera-se que este valor (a_{ij}) seja próximo de 0. A medida KMO varia entre 0 e 1 (quanto mais próxima da unidade, melhor é a adequabilidade da ACP aos dados em estudo) e, conforme sugerem Reis (2001) e Marôco (2014), deverá ser interpretado da seguinte forma: < 0,50 – inaceitável; 0,50 a 0,60 – má; 0,60 a 0,70 – razoável; 0,70 a 0,80 – média; 0,80 a 0,90 – boa; 0,90 a 1 – muito boa.

Poderá ainda ser utilizado o teste de esfericidade de Bartlett, que permite testar a hipótese de a matriz de correlações das variáveis de *input* na população ser uma matriz identidade. Interessa, portanto, rejeitar a hipótese em teste, já que, se a matriz de correlações das variáveis for uma matriz identidade, significa que elas não estão correlacionadas, o que invalida a aplicação da ACP aos dados em estudo.

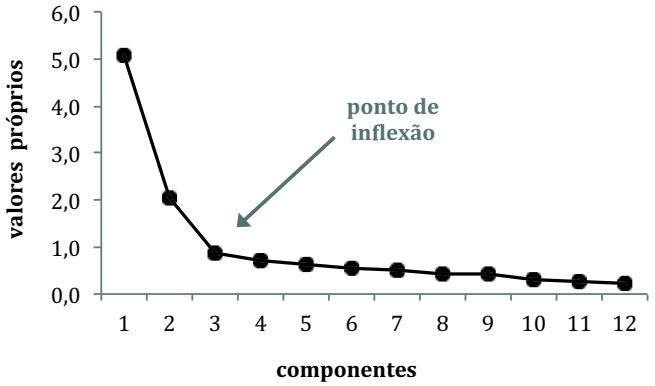
Retomando a nossa análise, podemos observar (Quadro 21) que temos indícios suficientes para considerar que este procedimento é adequado aos dados que estamos a trabalhar, já que KMO = 0,877 (adequabilidade boa), e que se rejeita a hipótese nula do teste de esfericidade de Bartlett de não existência de correlações significativas nas variáveis de *input* ($\chi^2_{(66)}$ = 8825,60, p < 0,001).

Quadro 21. Medidas de adequabilidade do procedimento aos dados

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,877
Bartlett's Test of Sphericity	Approx. Chi-Square	8825,600
	df	66
	Sig.	0,000

Como vimos, são calculadas tantas componentes quantas as variáveis de *input*. Contudo, esta não é uma solução desejável, já que o objectivo da ACP é a redução da informação. Existem vários critérios para decidir o número de componentes a reter. O critério de Kaiser, usado por omissão pelo SPSS, exclui as componentes que apresentem um valor próprio inferior a 1, ou seja, todas as componentes que tenham uma capacidade explicativa inferior à variância estandardizada de uma variável original (e que é igual a 1). O critério da variância explicada consiste em reter tantas componentes quantas sejam necessárias para explicar pelo menos 50% da variância total das variáveis de partida[16]. O critério do *scree plot* consiste na representação gráfica dos factores e dos valores próprios a eles associados, devendo reter-se as componentes até ao ponto de inflexão da curva. Neste caso, dever-se-iam reter, pelo menos, duas componentes.

Figura 8. Critério do *scree plot*



16 A percentagem de variância explicada é algo subjectiva e não existe um consenso acerca do valor mínimo. Marôco (2014) define 50% como mínimo aceitável; Reis (2001) sugere 70%.

No presente caso, o critério de Kaiser sugere a retenção de duas componentes, solução viável também se considerarmos o critério da percentagem mínima de variância explicada (Quadro 22).

Quadro 22. Variância explicada (ACP)

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5,069	42,244	42,244	5,069	42,244	42,244	4,062	33,849	33,849
2	2,041	17,012	59,256	2,041	17,012	59,256	3,049	25,407	59,256
3	,870	7,251	66,508						
4	,694	5,785	72,293						
5	,623	5,194	77,487						
6	,567	4,729	82,215						
7	,519	4,322	86,537						
8	,439	3,656	90,193						
9	,409	3,409	93,602						
10	,291	2,423	96,025						
11	,267	2,224	98,249						
12	,210	1,751	100,000						

Extraction Method: Principal Component Analysis.

Como observámos, os três critérios para a selecção do número de componentes a reter coincidem numa solução de duas componentes. Caso isto não se verificasse, deveríamos interpretar as diferentes soluções propostas: caso as diferentes soluções sejam igualmente interpretáveis, devemos optar pela solução com menor número de componentes (já que o objectivo da ACP

é a redução da informação); caso uma delas tenha um sentido substantivo mais claro, e faça mais sentido em termos interpretativos, deve ser essa a solução escolhida, mesmo que não seja a com menos componentes (equilíbrio entre ganhos em interpretabilidade e perdas em informação).

Tendo, então, optado por uma solução com duas componentes, podemos observar que esses dois factores explicam 59,23% da variância das variáveis iniciais. Importa agora proceder à interpretação destas duas componentes. A matriz das componentes dá-nos a informação relativamente à contribuição de cada variável para cada uma das componentes. Como vimos no início, as componentes são o resultado de combinações lineares de todas as variáveis, ou seja, todas as variáveis contribuem de alguma maneira para a formação de todas as componentes. Contudo, algumas variáveis contribuem mais do que outras, pelo que serão essas que diremos mais estruturadoras de cada uma das componentes. Dado que nem sempre a solução encontrada é facilmente interpretável (no caso de encontrarmos pesos factoriais elevados em mais do que uma componente, o que dificulta a percepção da componente para a qual a variável mais contribui), é comum adoptar-se um procedimento de rotação dos factores, que irá, portanto, melhorar a interpretabilidade e, logo, a utilidade científica do método (Tabachnick & Fidell, 2013). Os métodos de rotação têm, então, como objectivo a simplificação da estrutura factorial «dividindo o conjunto inicial de variáveis em subconjuntos tão independentes entre si quanto possível» (Reis, 2001, p. 275).

Vejamos graficamente um exemplo. Como se observa na Figura 9, as variáveis X_1, X_2 e X_3 parecem estar a distâncias relativamente semelhantes de ambas as componentes (tal como acontece com as variáveis X_4 e X_5), o que dificulta a decisão de para qual das componentes estão a contribuir mais. Aplicando um método de rotação (Figura 10), a solução torna-se mais clara, e conseguimos agora perceber, de forma mais evidente, que as variáveis X_1, X_2 e X_3 têm um maior peso na componente 2 e as variáveis X_4 e X_5 na componente 1. Note-se ainda que a estrutura inicial dos dados não sofreu qualquer alteração e que a proporção total de variância explicada não regista, igualmente, nenhuma mudança (Quadro 22).

Figura 9. Relação das variáveis com as componentes antes da rotação

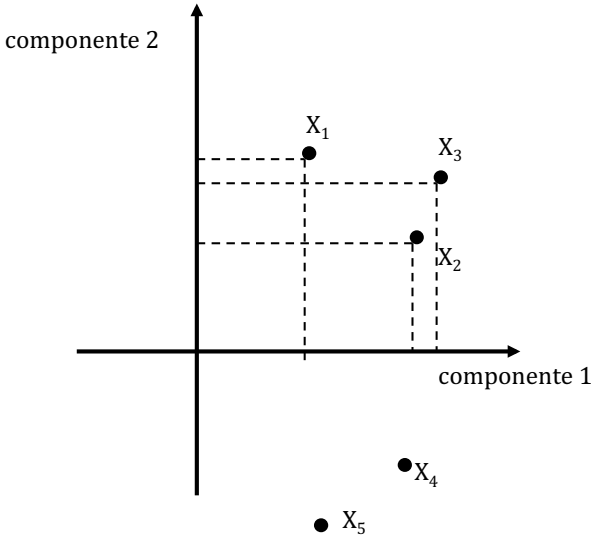
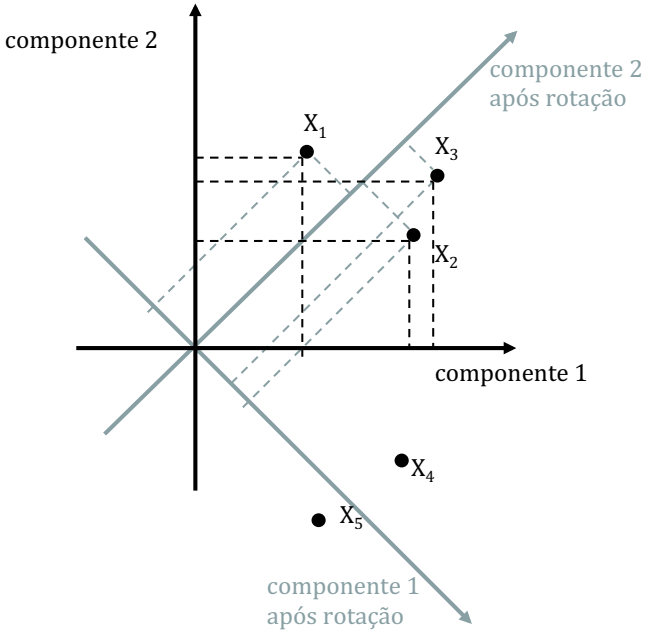


Figura 10. Relação das variáveis com as componentes antes e após rotação



Existem vários métodos de rotação disponíveis, mas a rotação VARIMAX é a mais comummente utilizada (Marôco, 2014; Reis, 2001; Tabachnick & Fidell, 2013). Através de um processo iterativo, maximiza a variação dos pesos factoriais em cada uma das componentes, de modo a que cada variável esteja sobretudo associada a apenas um dos factores. Será, portanto, este método que aplicaremos, nesta análise, para gerar a matriz das componentes que utilizaremos para a interpretação dos factores retidos (Quadro 23). Pode então perceber-se, através da análise das variáveis que mais contribuem[17] para a definição da primeira componente, que esta agrupa as questões relativas à cidadania e aos valores sociais e políticos (7 variáveis). Já a segunda componente tem subjacente o funcionamento das instituições democráticas, governo e partidos políticos (5 variáveis).

Quadro 23. Matriz dos pesos factoriais nas componentes rodadas

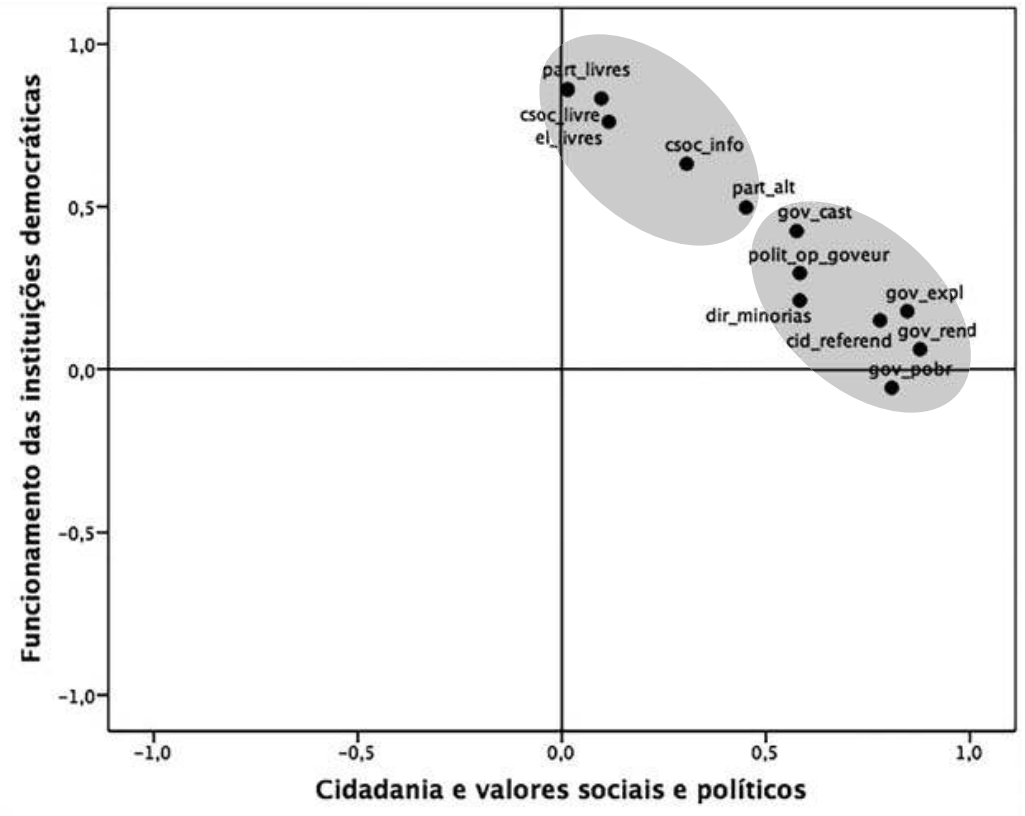
Rotated Component Matrix[a]		
	Component	
	1	2
Em Portugal, as eleições legislativas são livres e justas	,116	,763
Em Portugal, os diferentes partidos políticos apresentam alternativas claras entre si	,450	,502
Em Portugal, os partidos da oposição são livres para criticar o governo	,016	,861
Em Portugal, a comunicação social é livre para criticar o governo	,098	,834
Em Portugal, a comunicação social dá aos cidadãos informação correcta para avaliar o governo	,305	,634
Em Portugal, os direitos das minorias são protegidos	,583	,213
Em Portugal, os cidadãos têm a última palavra nos assuntos políticos mais importantes votando diretamente sobre eles em referendos	,778	,152
Em Portugal, os partidos do governo que fazem um mau trabalho são castigados nas eleições	,575	,425
Em Portugal, o governo protege todos os cidadãos da pobreza	,806	-,055
Em Portugal, o governo explica as suas decisões aos eleitores	,844	,180
Em Portugal, o governo toma medidas para reduzir as diferenças nos níveis de rendimento	,875	,063
Em Portugal, os políticos têm em conta as opiniões de outros governos Europeus antes de tomarem decisões	,583	,297

Extraction Method: Principal Component Analysis.
a. Rotation converged in 3 iterations.

17 As variáveis que se consideram como tendo maior contribuição para a definição da componente são aquelas que apresentam um peso factorial superior a 0,5.

Se quisermos analisar as projecções factoriais em dois planos ortogonais, verificamos a forte correlação positiva entre as variáveis de cada um dos grupos: o das que privilegiam a dimensão da cidadania e o das que remetem para o funcionamento das instituições governamentais e partidos políticos (Figura 11).

Figura 11. Projecções factoriais das componentes



Legenda: el_livres: Em Portugal, as eleições legislativas são livres e justas; part_alt: Em Portugal, os diferentes partidos políticos apresentam alternativas claras entre si; part_livres; Em Portugal, os partidos da oposição são livres para criticar o governo; csoc_livre: Em Portugal, a comunicação social é livre para criticar o governo; csoc_info: Em Portugal, a comunicação social dá aos cidadãos informação correcta para avaliar o governo; dir_minorias: Em Portugal, os direitos das minorias são protegidos; cid_referend: Em Portugal, os cidadãos têm a última palavra nos assuntos políticos mais importantes votando diretamente sobre eles em referendos; gov_cast: Em Portugal, os partidos do governo que fazem um mau trabalho são castigados nas eleições; gov_pobr: Em Portugal, o governo protege todos os cidadãos da pobreza; gov_expl: Em Portugal, o governo explica as suas decisões aos eleitores; gov_rend: Em Portugal, o governo toma medidas para reduzir as diferenças nos níveis de rendimento; polit_op_goveur: Em Portugal, os políticos têm em conta as opiniões de outros governos Europeus antes de tomarem decisões.

Após a interpretação das componentes, elas podem ser operacionalizadas na base de dados, através da gravação dos *scores* factoriais, que passam, então, a fazer parte da base de dados, e a poder ser utilizadas como quaisquer outras variáveis. É apenas de referir que as novas variáveis são estandardizadas, pelo que apresentam uma média igual a 0 e um desvio-padrão igual a 1. Apesar de perder-se a escala inicial, os limites entre os quais estas variáveis variam correspondem, substantivamente, aos pólos das variáveis originais. Neste caso, e para dar um exemplo, valores elevados na variável a que poderíamos chamar de «adequado funcionamento das instituições democráticas, governo e partidos políticos» (e que corresponde à componente 2) significam uma elevada concordância, já que, relembremos, as variáveis originais foram medidas numa escala entre 0 = «não se aplica nada» e 10 = «aplica-se totalmente».

Poderíamos, a partir daqui, e retomando o problema inicial, explorar se homens e mulheres revelam ou não percepções diferenciadas relativamente ao funcionamento do sistema político, nomeadamente através da aplicação de um teste *t* à comparação de médias entre dois grupos (caso se cumpram os pressupostos de aplicação).

Análise de Correspondências Múltiplas (ACM)

O aparecimento da ACM dá-se pela mão de Benzécri (1976), cujos trabalhos vieram contribuir, de forma significativa, para o desenvolvimento de técnicas de análise adequadas ao tipo de variáveis mais frequentemente usadas na área da Sociologia (categóricas). De facto, até aí, o único método quantitativo disponível para o tratamento de variáveis nominais era o teste do Qui^2, apresentando a limitação de apenas poder aferir a independência entre as duas variáveis de partida (Lisboa, 2014). Como nos diz Henrique Garcia Pereira, «a Análise das Correspondências, mesmo na sua versão inicial, permitia estabelecer relações no interior de cada conjunto de modalidades (e entre os dois conjuntos), abrindo assim a porta para a possibilidade de uma certa modelização de variáveis qualitativas, o que constitui um importantíssimo avanço

no processamento estatístico de tais variáveis» (2008: 9). Para além da Análise das Correspondências de Benzécri (escola francesa), foram sendo desenvolvidos outros métodos, com o mesmo propósito, pela escola americana (ligada à escola de Leiden). Tal como descreve Garcia Pereira (2008), a disputa entre estas duas correntes parece já ter sido resolvida «com elegância» por Helena Carvalho (2008), cujo trabalho veio demostrar a equivalência entre as duas abordagens em competição.

A Análise de Correspondências Múltiplas (ACM) constitui-se como uma técnica descritiva de redução da informação a um pequeno número de dimensões que expliquem a estrutura subjacente das relações que se estabelecem entre objectos (indivíduos) e categorias, conferindo-lhe maior interpretabilidade (Carvalho, 2008; Hair *et al.*, 2013). Esta análise apresenta enormes potencialidades no contexto das Ciências Sociais, já que, por um lado, trabalha com variáveis categóricas (nominais e também ordinais, ainda que a ordem não seja tida em consideração), as mais comuns nesta área; e, por outro lado, possibilita a apresentação dos resultados graficamente, num mapa perceptual[18], o que facilita grandemente a comunicação desses mesmos resultados.

Uma explicação mais detalhada dos fundamentos estatísticos e dos procedimentos em SPSS pode ser consultada em Carvalho (2008), pelo que aqui apresentaremos um exemplo de aplicação da ACM na área das desigualdades sociais. Para este exemplo, mobilizaremos um conjunto diferente de dados, relativo a um inquérito sociológico realizado em 2007, com vista à caracterização das actividades desenvolvidas pelas crianças e jovens em Portugal, e cujos principais resultados foram já publicados por Lisboa *et al.* (2009). Este trabalho teve como um dos seus principais objectivos o estudo dos factores associados à produção e reprodução do fenómeno do trabalho infantil em Portugal. Para tal, foi realizado um inquérito, que incluiu crianças e jovens que frequentassem

18 Um mapa perceptual corresponde a uma «Visual representation of a respondent's perceptions of objects on two or more dimensions. Usually this map has opposite levels of dimensions on the ends of the X and Y axes, such as "sweet" to "sour" on the ends of the X axis and "high-priced" to "low-priced" on the ends of the Y axis. Each object then has a spatial position on the perceptual map that reflects the relative similarity or preference to other objects with regard to the dimensions of the perceptual map» (Hair *et al.*, 2013: 520).

escolas onde decorriam os programas PIEF[19]. Recorrendo a duas amostras emparelhadas (uma de alunos/as a frequentar o ensino regular e outra de alunos/as a frequentar o programa PIEF), tentou-se perceber se existiam ou não perfis diferenciados destes/as alunos/as relativamente aos meios de socialização, às expectativas, ao aproveitamento escolar e ao relacionamento com a escola, e de que forma eles se poderiam apresentar como condicionantes da emergência ou manutenção de situações de trabalho infantil. Pretendendo perceber como se estrutura esse espaço de condicionantes e práticas escolares, recorreu-se a uma ACM, cujos resultados foram já publicados (Lisboa & Malta, 2009). Pretendemos aqui retomar essa mesma análise, acrescentando-lhe apenas um maior detalhe na descrição da aplicação do procedimento, algo que não caberia no âmbito da publicação referida. O espaço de condicionantes e práticas escolares foi estruturado pelas seguintes variáveis, com as correspondentes categorias (Quadro 24):

Quadro 24. Variáveis e categorias mobilizadas para a construção do espaço de condicionantes e práticas escolares

Variável	Categorias	Rótulos usados na ACM
Expectativas de escolaridade	9º ano	exp_9ºano
	12º ano	exp_12ºano
	Universidade	exp_univ
Retenção escolar	Sim	R_S
	Não	R_N
Qualificações da mãe	1º ciclo	Q_M_1º ciclo
	2º ciclo	Q_M_2º ciclo
	3º ciclo	Q_M_3º ciclo
	Secundário/superior	Q_M_Sec_Sup

19 O Programa Integrado de Educação e Formação (PIEF), implementado pelo Plano para a Prevenção e Eliminação da Exploração do Trabalho Infantil (PETI), na dependência do Ministério do Trabalho e da Solidariedade Social, constitui-se como um instrumento de combate a situações, efectivas ou emergentes, de trabalho infantil e abandono escolar.

Variável	Categorias	Rótulos usados na ACM
Qualificações do pai	1º ciclo	Q_P_1º ciclo
	2º ciclo	Q_P_2º ciclo
	3º ciclo	Q_P_3º ciclo
	Secundário/superior	Q_P_Sec_Sup
Tem computador	Sim	comp_S
	Não	comp_N
Tem internet	Sim	internet_S
	Não	internet_S
Abandono parcial (desistências)	Sim	D_S
	Não	D_N
Currículo escolar frequentado	PIEF	PIEF
	Ensino regular	ER
Trabalho infantil no passado*	Sim	TI_S
	Não	TI_N

* Varialvel suplementar.

Em primeiro lugar, há que perceber quantas dimensões principais estão subjacentes à estrutura dos dados que estamos a trabalhar. A selecção do número de dimensões a reter é feita com base na quantificação da variância explicada por cada dimensão (valor próprio) ou, preferencialmente, pela inércia de cada dimensão, já que esta dá conta da variância explicada em termos relativos (e é dada pela divisão entre o valor próprio e o número de variáveis activas[20]). Esta variância pode ser entendida como a capacidade de cada uma das dimensões explicar a relação entre os dados de origem (a sua variabilidade). A inércia varia entre 0 e 1, e

20 O procedimento da ACM permite a inclusão de variáveis com diferentes estatutos: as que são mobilizadas para a estruturação do espaço são as variáveis activas; as variáveis definidas como suplementares são integradas na análise, não para estruturar o espaço, mas apenas para perceber a relação que estabelecem, tanto com as variáveis (e categorias) activas como com as dimensões definidas.

quanto mais perto do limite superior, mais variância é explicada pela dimensão. A inércia é decrescente, *i. e.*, a primeira dimensão regista o maior valor. As dimensões mais relevantes são as que tiverem associados valores de inércia mais elevados, e devemos reter as dimensões que se situem antes de descidas acentuadas no decréscimo da inércia (Carvalho, 2008).

Para avaliar o decréscimo da inércia e decidir quanto ao número de dimensões a reter, é necessário realizar uma ACM com o total de dimensões que é possível obter com os dados de origem. Quando não existem não respostas, o número do total de dimensões é dado por *p-m*, sendo *p* o número total de categorias das variáveis activas e *m* o número de variáveis activas sem casos omissos. Se todas as variáveis tiverem casos omissos, o número máximo de dimensões é dado por *p*-1. No caso em estudo, o número total de categorias em análise é de 21 e todas as 8 variáveis têm casos omissos (Quadro 25).

Quadro 25. Casos válidos e omissos nas variáveis consideradas para a ACM

Statistics									
		Expectativas de escolaridade	Retenção escolar	Abandono parcial (desistências)	Qualificações da mãe	Qualificações do pai	Tem computador	Tem internet	Currículo escolar frequentado
N	Valid	1288	1368	1223	1115	873	1368	1368	1368
	Missing	80	0	145	253	495	0	0	0

Assim, o número máximo de dimensões nesta ACM é de 21 – 4 = 17, como se observa no quadro seguinte (Quadro 26).

Quadro 26. Distribuição dos valores próprios e da inércia para o espaço de condicionantes e práticas escolares

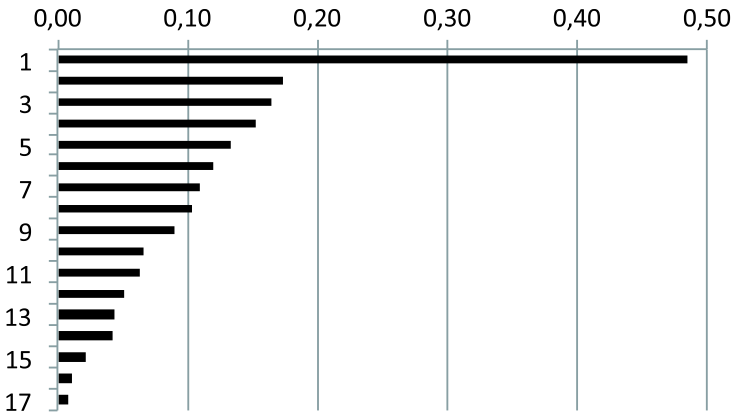
Model Summary			
Dimension	Cronbach's Alpha	Variance Accounted For	
		Total (Eigenvalue)	Inertia
1	,848	3,874	,484
2	,321	1,391	,174
3	,276	1,319	,165
4	,205	1,218	,152
5	,075	1,071	,134
6	-,048	,959	,120
7	-,154	,881	,110
8	-,229	,833	,104
9	-,444	,720	,090
10	-,990	,536	,067
11	-1,108	,508	,063
12	-1,603	,416	,052
13	-2,109	,351	,044
14	-2,175	,344	,043
15	-5,529	,171	,021
16	-11,237	,092	,012
17	-17,840	,060	,008
Total		14,746	1,843
Mean	-,175[a]	,867	,108

a. Mean Cronbach's Alpha is based on the mean

Analisando graficamente o decréscimo da inércia (Figura 12), percebemos que a variância relativa explicada se reduz bastante a partir da segunda dimensão. Apesar da segunda dimensão apresentar uma capacidade explicativa bastante reduzida relativa-

mente à primeira dimensão, poderemos reter as duas primeiras, considerando a sua interpretabilidade, como veremos de seguida.

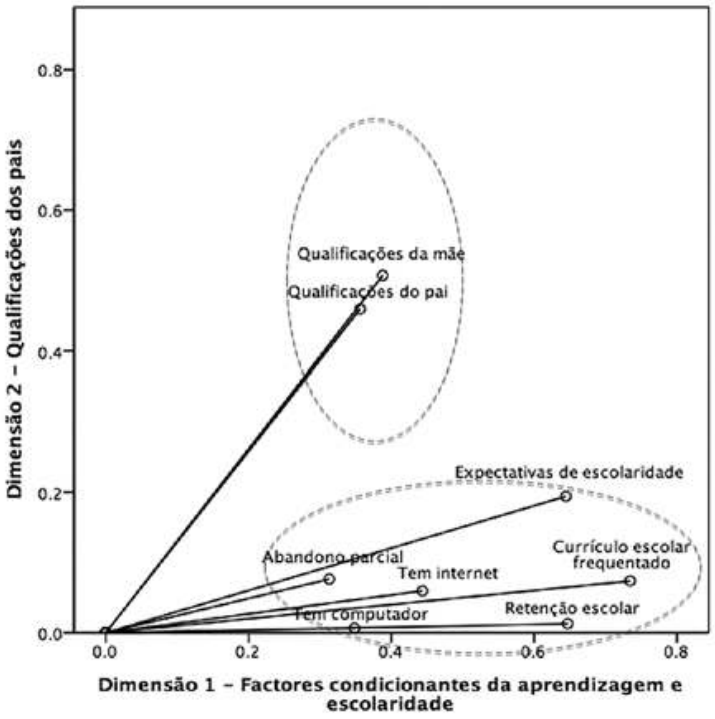
Figura 12. Representação gráfica do decréscimo da inércia para o espaço de condicionantes e práticas escolares



Para obter a solução da ACM com duas dimensões, é necessário voltar a fazer o procedimento, especificando agora o número de dimensões a reter (duas). De forma a identificar substantivamente as duas dimensões que estão subjacentes à estrutura dos dados, é necessário perceber que variáveis mais contribuem para a estruturação do espaço e em que dimensão mais contribuem para discriminar os indivíduos.

A análise da figura seguinte (Figura 13), conjuntamente com a leitura da tabela relativa às medidas de discriminação (Quadro 27), indica-nos que a dimensão 1 será estruturada por variáveis que remetem para factores que condicionam os processos escolares. De facto, aquelas que mais contribuem para discriminar os indivíduos estão relacionadas com o sucesso escolar (retenção), com as expectativas de escolaridade, com o relacionamento com a escola (desistências e currículo escolar frequentado) e com o acesso à informação (se tem ou não computador e Internet em casa). Por outro lado, a segunda dimensão remete claramente para a qualificação dos pais, importantes para estrutura de oportunidades destes/as alunos/as.

Figura 13. Medidas de discriminação das variáveis nas duas dimensões retidas



Quadro 27. Medidas de discriminação das variáveis nas duas dimensões retidas

Discrimination Measures			
	Dimension		
	1	2	Mean
Expectativas de escolaridade	,644	,194	,419
Retenção escolar	,647	,013	,330
Abandono parcial	,313	,076	,195
Qualificações da mãe	,388	,507	,448
Qualificações do pai	,356	,460	,408
Tem computador	,348	,006	,177
Tem internet	,443	,060	,252
Currículo escolar frequentado	,734	,074	,404
Active Total	3,874	1,390	2,632

Importa agora ver, para cada variável, quais são as categorias que mais discriminam os indivíduos e que mais vão contribuir para a estruturação do espaço e dos perfis dos/as alunos/as relativamente às condicionantes e práticas escolares. E isto é feito com base na contribuição que cada categoria tem para a variância relativa de cada dimensão ou factor. A soma das contribuições de todas as categorias para cada dimensão é igual a 1. Assim, toma-se a média das contribuições como valor a partir do qual a categoria deve ser considerada pertinente, média essa que é dada por $\frac{1}{n.º \text{ de categorias das variáveis activas}}$. Neste caso, e considerando o número de categorias presentes (Quadro 24), a contribuição média é de 1 / 21 = 0,048. A partir deste valor, podemos identificar as categorias que mais contribuem para discriminar os indivíduos e verificar qual é o seu posicionamento no mapa perceptual. Vejamos, então, como se processa esta análise.

Tomemos a variável «Tem Internet». Como se observa no quadro seguinte (Quadro 28), a categoria «Internet_S» apresenta uma contribuição (0,067) acima da contribuição média (0,048) das categorias para a variância relativa da dimensão 1 e a categoria «Internet_N» contribui para a discriminação dos indivíduos também na dimensão 1 (com uma contribuição acima da média, de 0,048).

Quadro 28. Contribuições das categorias da variável «Tem internet»

Tem internet								
Points: Contributions								
				Contribution				
				Of Point to Inertia of		Of Dimension to Inertia of Point		
Category	Frequency	Mass	Inertia	1	2	1	2	Total
Internet_S	491	,049	,088	,067	,031	,367	,061	,428
Internet_N	877	,088	,053	,048	,012	,439	,039	,478
Active Total		,137	,141	,114	,043			

Variable Principal Normalization.

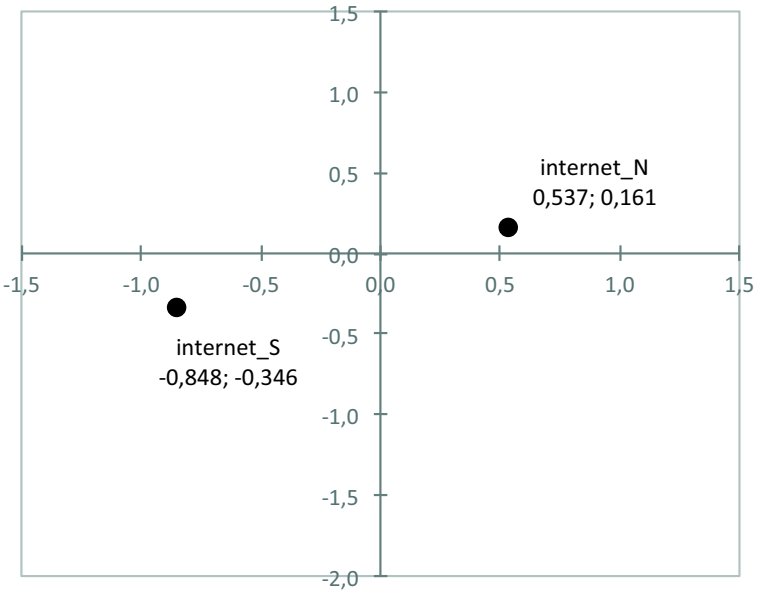
Perante essas contribuições significativas, podemos então ver qual é o posicionamento no espaço dessas mesmas categorias, quer através das suas coordenadas nas dimensões 1 e 2 (Quadro 29), quer através da sua representação gráfica ao serem projectadas no plano (Figura 14).

Quadro 29. Coordenadas das categorias da variável «Tem Internet»

Tem internet			
Points: Coordinates			
		Centroid Coordinates	
		Dimension	
Category	**Frequency**	**1**	**2**
Internet_S	491	-,848	-,346
Internet_N	877	,537	,161

Variable Principal Normalization.

Figura 14. Projecção das categorias da variável «Tem Internet» no espaço



Olhemos para outro exemplo, agora relativo ao nível de escolaridade do pai. Como se observa na tabela seguinte (Quadro 30), a categoria «Q_P_Sec_Sup» contribui para discriminar os indivíduos na dimensão 1 (0,050>0,048) e na dimensão 2 (0,193 > 0,048); já a categoria «Q_P_2 ciclo» discrimina os indivíduos na dimensão 2 (0,114 > 0,048).

Quadro 30. Contribuições das categorias da variável «Qualificações do pai»

Qualificações do pai

Points: Contributions

				Contribution				
				Of Point to Inertia of		Of Dimension to Inertia of Point		
Category	**Frequency**	**Mass**	**Inertia**	**1**	**2**	**1**	**2**	**Total**
Q_P_1 ciclo	421	,042	,088	,023	,004	,128	,008	,136
Q_P_2 ciclo	154	,015	,113	,001	,114	,005	,176	,181
Q_P_3 ciclo	156	,016	,112	,017	,019	,075	,030	,105
Q_P_Sec_Sup	142	,014	,114	,050	,193	,214	,295	,509
Missing	495							
Active Total		,088	,426	,092	,331			

Variable Principal Normalization.

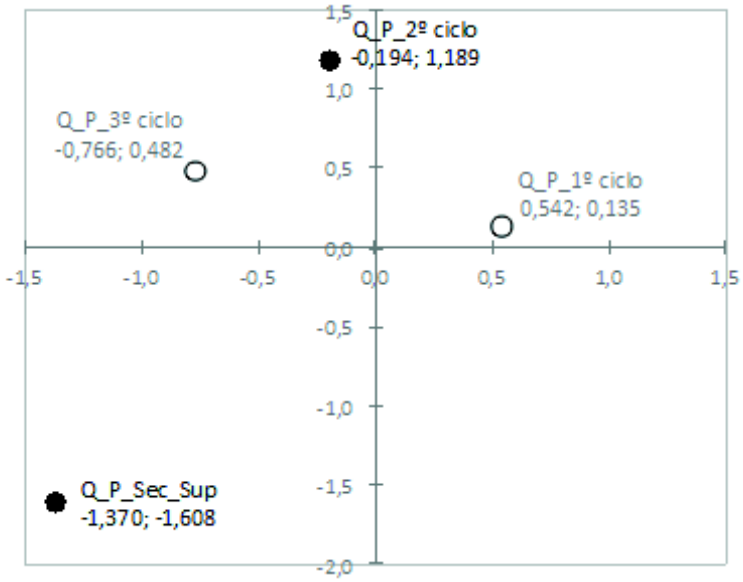
Podemos então perceber qual é o posicionamento destas categorias no plano (Quadro 31) e fazer a sua projecção no espaço (Figura 15). Note-se que fizemos uma distinção entre as categorias que discriminam os indivíduos (numa ou noutra dimensão) e as que não contribuem de forma significativa, correspondendo o círculo preto às primeiras e o círculo branco às segundas. As categorias que menos discriminam os indivíduos devem ser mantidas na representação, uma vez que contribuem, de facto, para a estruturação do espaço.

Quadro 31. Coordenadas das categorias da variável «Qualificações do pai»

Qualificações do pai			
Points: Coordinates			
		Centroid Coordinates	
		Dimension	
Category	Frequency	1	2
Q_P_1 ciclo	421	,542	,135
Q_P_2 ciclo	154	-,194	1,189
Q_P_3 ciclo	156	-,766	,482
Q_P_Sec_Sup	142	-1,370-	1,608
Missing	495		

Variable Principal Normalization.

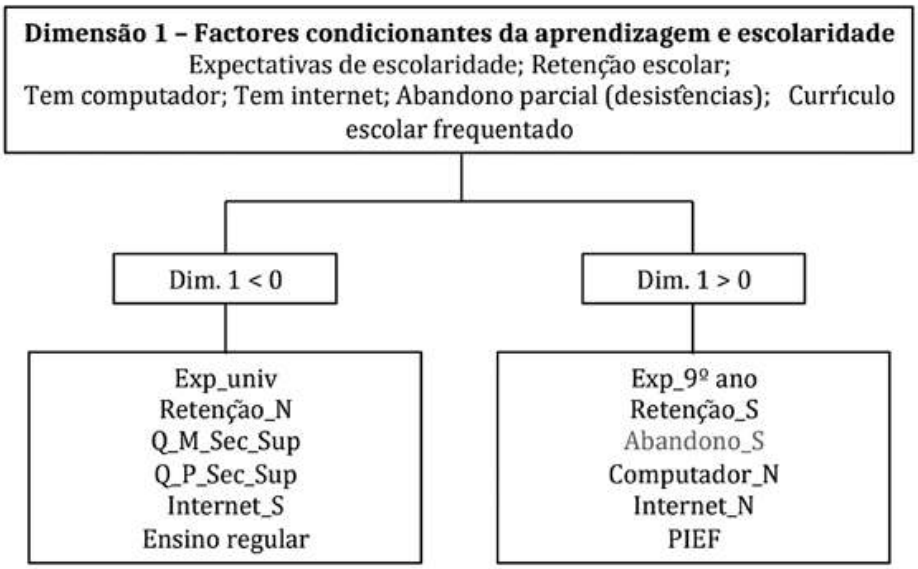
Figura 15. Projecção das categorias da variável «Qualificações do pai» no espaço



De forma semelhante, analisamos as restantes variáveis e respectivas categorias. Note-se que a leitura e a interpretação dos dados deve sempre partir das tabelas das quantificações e das coordenadas, e nunca da representação gráfica relativa à projecção das categorias das variáveis no espaço. Deveremos então proceder a uma esquematização desta informação, para que possamos compreender melhor a estrutura das dimensões de análise.

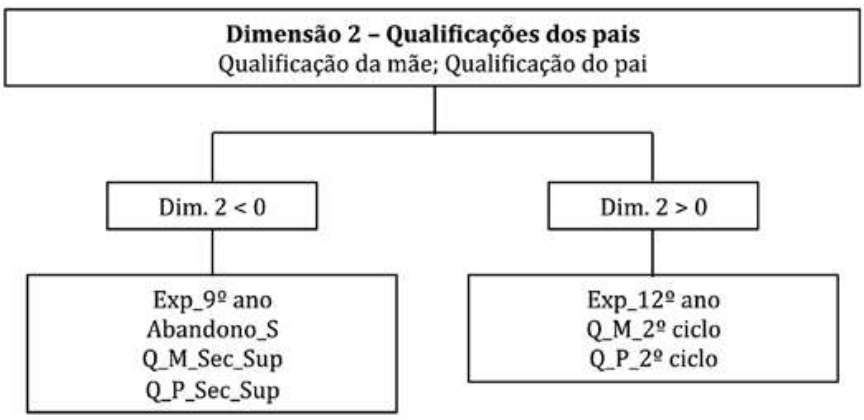
Podemos então perceber que a dimensão 1, relativa aos factores condicionantes da aprendizagem e escolaridade, varia num eixo que vai de um ambiente mais desfavorável a um ambiente mais favorável (Figura 16). Por outro lado, a dimensão 2, relativa às qualificações dos pais, varia num eixo que vai das maiores qualificações a menores qualificações escolares (Figura 17).

Figura 16. Descrição da dimensão 1



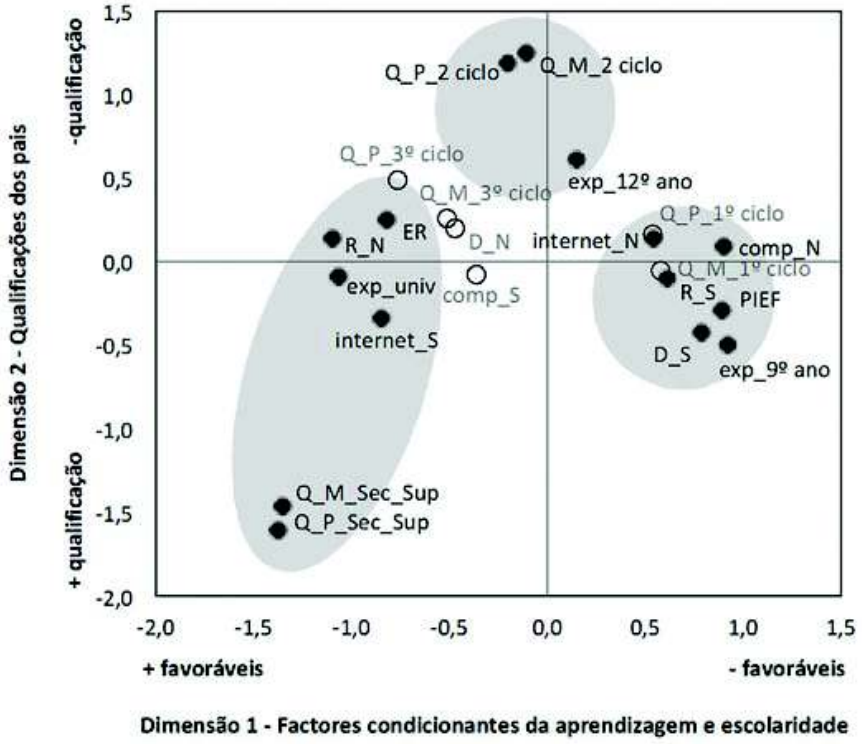
Nota: A categoria a cinzento apresenta uma contribuição abaixo mas muito próxima da contribuição média.

Figura 17. Descrição da dimensão 2



Agregando então toda esta informação, chegamos à análise dos perfis de condicionantes e práticas escolares (Figura 18).

Figura 18. Espaço de condicionantes e práticas escolares

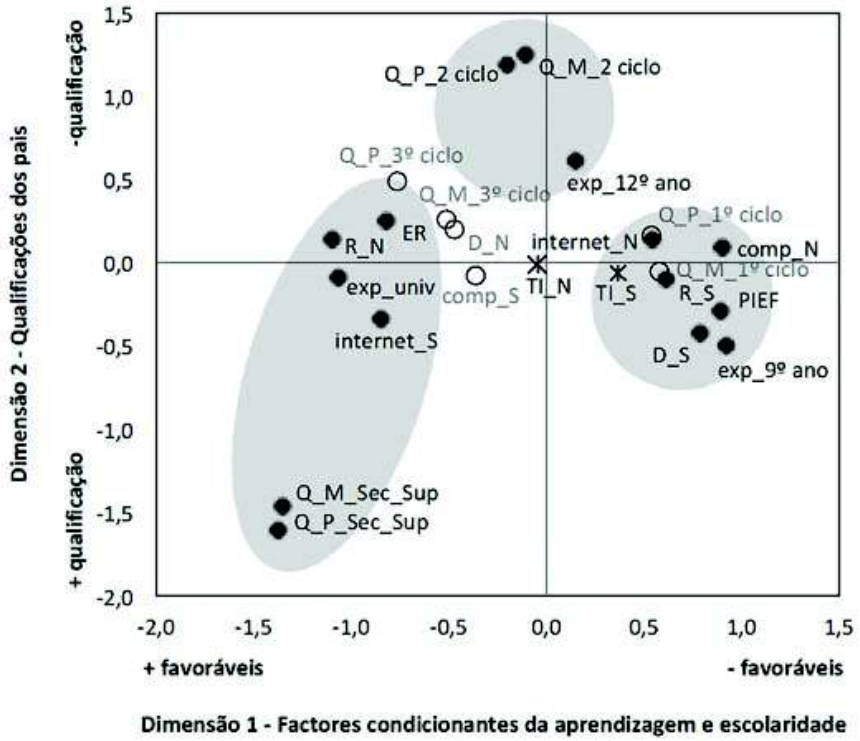


Como se observa no mapa perceptual (Figura 18), podemos identificar três perfis diferenciados. No primeiro perfil (que se localiza nos quadrantes 3 e 4), encontramos alunos/as que frequentam o ensino regular, que têm expectativas elevadas de prosseguimento dos estudos (esperam prosseguir para o ensino superior), que nunca ficam retidos/as, que têm Internet em casa e cujos pais detêm capitais escolares elevados (ao nível do ensino secundário e da licenciatura). O segundo perfil (situado nos 1.º e 2.º quadrantes) inclui alunos/as que demonstram uma relação mais problemática com a escola (com experiência(s) de retenção e desistência), que não têm computador nem Internet em casa, que têm baixas expectativas de prosseguimento dos estudos (apenas ao nível da escolaridade obrigatória à época – 9.º ano) e que frequentam o

PIEF. Já o terceiro perfil encontrado (situado nos quadrantes 1 e 4) apresenta menos especificidades relativamente ao conjunto das variáveis consideradas, associando qualificações dos pais relativamente baixas (2.º ciclo) a expectativas intermédias de prosseguimento dos estudos (até ao 12.º ano).

O procedimento da ACM permite ainda incluir variáveis suplementares na análise que, não estruturando o espaço, vão ser projectadas nele, para que possamos perceber a relação que estabelecem com o espaço já definido (tanto em termos de categorias activas, como de dimensões e de perfis). Nesse sentido, porque o espaço de condicionantes e práticas escolares foi construído com o objectivo de perceber quais seriam os factores com mais peso no percurso dos/as jovens relativamente à emergência e manutenção de situações de trabalho infantil, foi projectada no plano a variável «Trabalho infantil no passado». Como se observa (Figura 19), a estrutura de relações manteve-se inalterada, já que as variáveis que estruturam o espaço (variáveis activas) se mantiveram as mesmas.

Figura 19. Espaço de condicionantes e práticas escolares com projecção de «Trabalho infantil no passado» como variável suplementar



Através da projecção desta variável suplementar (Figura 19), percebe-se que os/as alunos/as com experiência de trabalho infantil no passado estão associados/as ao segundo perfil, ou seja, ao conjunto de jovens que revelam um percurso escolar mais instável e um enquadramento familiar mais desfavorável. Já os/as alunos/as que não tiveram experiências de trabalho infantil no passado não surgem associados a nenhum dos perfis definidos. Encontrada esta topologia, podemos operacionalizar estes perfis numa tipologia por forma a, recorrendo a outro tipo de instrumentos, conhecer melhor os grupos aqui definidos. Para tal, e uma vez que a ACM não é um método de agrupamento, devemos recorrer a uma análise de *clusters*, que permitirá criar uma nova variável que congregue as unidades de análise (neste caso, os/as alunos/as) em diferentes grupos, de acordo com os seus perfis diferenciados, exercício que faremos de seguida.

Análise de *Clusters*

Tal como referido anteriormente, podemos operacionalizar os perfis encontrados a partir da ACM. Em todo o caso, é de notar que a análise de *clusters* é aplicável a uma diversidade de outras situações, desde que o objectivo seja o de agrupar elementos segundo a sua (dis)semelhança[21]. A análise de *clusters* constitui-se como um método de análise de dados multivariada, cujo objectivo é o de encontrar agrupamentos «naturais» de indivíduos: «This is done by grouping individuals that are "similar" according to some appropriate criterion» (Härdle & Simar, 2007: 274). Apesar de ser tipicamente associada a variáveis métricas, a análise de *clusters* pode ser realizada com variáveis categóricas (sejam nominais ou ordinais), desde que seleccionadas medidas de semelhança aplicáveis a esse tipo de variáveis (Everitt, Landau, Leese & Stahl, 2011; Hair *et al.*, 2013; Jain, Murty & Flynn, 1999).

Considerando que pretendemos operacionalizar os três perfis encontrados, sabemos, à partida, o número de *clusters* que

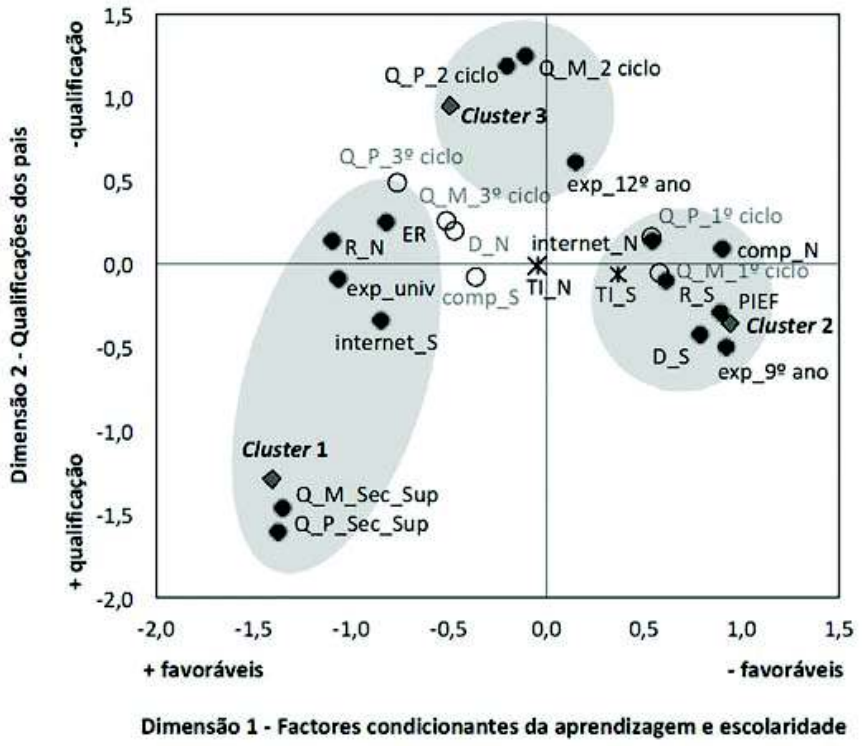
21 Para outras aplicações da análise de *clusters*, ver, e.g., Marôco (2014).

vamos obter, pelo que optamos pelo método de agrupamento *k-means,* que se constitui como um método de optimização que se baseia numa definição *a priori* do número de grupos que conterão todos os indivíduos, de modo a que «dentro de cada grupo os elementos sejam o mais semelhante possível e o mais diferente possível de elementos de outros grupos» (Reis, 2001: 296). Se não fosse este o caso (se não soubéssemos à partida o número de grupos a constituir), teríamos de recorrer a métodos de agrupamento hierárquicos que «recorrem a passos sucessivos de agregação dos sujeitos considerados individualmente, *i.e.* cada sujeito é um *Cluster*, e depois estes vão sendo agrupados de acordo com as suas proximidades (neste caso os métodos dizem-se aglomerantes), ou pelo contrário todos os sujeitos são, à partida, agrupados num único *Cluster* e depois são divididos em subgrupos de acordo com as suas medidas de distância (neste caso os métodos dizem-se divisivos)» (Marôco, 2014: 554). Cabe então ao/à investigador/a decidir quantos *clusters* reter, baseando-se na leitura das distâncias entre *clusters* (disponível através dos coeficientes apresentados pelo SPSS na tabela *Aglomeration schedule*)[22].

Para esta análise, vamos recorrer aos *scores* dos indivíduos nas dimensões 1 e 2 (gravados através do procedimento da ACM), que correspondem às suas coordenadas no plano. Gravado na base de dados o grupo de pertença de cada um dos indivíduos (resultante da aplicação do procedimento *k-means*), podemos projectar no espaço gerado pela ACM os centróides de cada um dos *clusters*, por forma a validar os resultados obtidos. Como se observa, os *clusters* ocupam posições próximas aos perfis que identificámos via ACM (Figura 20).

22 Tendo em conta que ambos os métodos hierárquicos e não hierárquicos apresentam vantagens e desvantagens, deve usar-se uma combinação dos dois, ou seja, aplicar em primeiro lugar um método hierárquico e, definido o número de *clusters*, operacionalizá-los através de um método não hierárquico (Hair *et al.*, 2013; Marôco, 2014).

Figura 20. Disposição dos *clusters* no espaço de condicionantes e práticas escolares



Podemos então fazer novos cruzamentos, a partir da pertença aos *clusters*, de modo, não apenas a confirmar a configuração dos perfis, mas também para caracterizá-los mais aprofundadamente. Através de uma tabela cruzada com os resíduos estandardizados ajustados (já descritos anteriormente), é possível descrever os grupos encontrados. Como se observa (Quadro 32), as características de cada um dos grupos são muito semelhantes às descritas anteriormente, aquando da análise dos perfis encontrados na ACM. E percebemos ainda que, à parte do nível de instrução dos pais, os *clusters* 1 e 3 são muito semelhantes. O que os distingue é, efectivamente, a escolaridade dos pais e diferentes graus de associação relativamente às restantes variáveis (não se encontrando grandes diferenças no que diz respeito à retenção escolar, ao ter computador e ao trabalho infantil no passado).

Quadro 32. Cruzamento da pertença aos *clusters* com as variáveis estruturadoras do espaço de condicionantes e práticas escolares

Variável	Categoria		Cluster 1	Cluster 2	Cluster 3
Expectativas de escolaridade	exp_9º ano	Count	10	411	29
		Adjusted Residual	-9,9	24,0	-17,2
	exp_12 ano	Count	29	154	228
		Adjusted Residual	-6,0	-4,1	8,8
	exp_univ	Count	168	25	234
		Adjusted Residual	16,0	-20,3	8,7
Retenção escolar	R_S	Count	49	630	231
		Adjusted Residual	-14,9	23,0	-12,5
	R_N	Count	167	16	275
		Adjusted Residual	14,9	-23,0	12,5
Abandono parcial (desistências)	D_S	Count	29	318	46
		Adjusted Residual	-6,4	18,5	-13,8
	D_N	Count	184	206	440
		Adjusted Residual	6,4	-18,5	13,8
Qualificações da mãe	Q_M_1 ciclo	Count	13	359	158
		Adjusted Residual	-12,1	16,7	-7,6
	Q_M_2 ciclo	Count	5	48	157
		Adjusted Residual	-6,1	-6,2	10,8
	Q_M_3 ciclo	Count	25	48	140
		Adjusted Residual	-2,1	-6,4	8,0
	Q_M_Sec_Sup	Count	142	12	8
		Adjusted Residual	26,3	-9,6	-10,2
Qualificações do pai	Q_P_1 ciclo	Count	4	264	153
		Adjusted Residual	-12,9	14,9	-4,4
	Q_P_2 ciclo	Count	5	33	116
		Adjusted Residual	-5,4	-4,5	8,6
	Q_P_3 ciclo	Count	24	25	107
		Adjusted Residual	-1,1	-6,1	6,8
	Q_P_Sec_Sup	Count	129	5	8
		Adjusted Residual	24,2	-9,1	-10,1
Tem computador	comp_S	Count	205	321	405
		Adjusted Residual	9,2	-13,8	7,3
	comp_N	Count	11	325	101
		Adjusted Residual	-9,2	13,8	-7,3
Tem internet	Internet_S	Count	181	106	204
		Adjusted Residual	16,0	-14,2	2,6
	Internet_N	Count	35	540	302
		Adjusted Residual	-16,0	14,2	-2,6
Currículo escolar frequentado	PIEF	Count	18	604	62
		Adjusted Residual	-13,3	30,4	-21,4
	ER	Count	198	42	444
		Adjusted Residual	13,3	-30,4	21,4
Trabalho infantil no passado	TI_S	Count	26	177	83
		Adjusted Residual	-3,5	5,6	-3,1
	TI_N	Count	190	469	423
		Adjusted Residual	3,5	-5,6	3,1

Assim, e tal como descrito por Lisboa e Malta, «o primeiro [perfil] engloba genericamente alunos com um enquadramento cultural, económico e familiar favorável, com níveis de sucesso escolar altos; o segundo [que corresponde aqui ao *cluster* 3] reúne alunos com um enquadramento cultural, económico e familiar misto, com sucesso escolar irregular; e o terceiro [*cluster* 2] concentra os alunos com piores níveis de sucesso escolar, cujo enquadramento cultural, económico e familiar é desfavorável» (2009: 110).

É, como se disse, ainda possível explorar associações adicionais que ajudem a caracterizar, de forma mais detalhada, as circunstâncias de aprendizagem e práticas escolares dos/as alunos/as de cada um dos grupos. Como se observa no quadro seguinte (Quadro 33), os/as jovens que pertencem ao perfil com um enquadramento mais desfavorável são também aqueles/as que já tiverem um processo disciplinar e que consideram mau o seu desempenho escolar; relativamente às práticas escolares e condições de aprendizagem, esses/as são também os/as alunos/as que não têm um quarto só para si em casa, que dizem não ter ou nunca fazer os trabalhos de casa, e que se deitam mais tarde em dias de escola (afirmando não ter horas para deitar-se ou deitar-se depois da meia-noite). Relativamente aos perfis 1 e 3, estes são sobretudo alunos/as que nunca tiveram um processo disciplinar, que têm um quarto só para si em casa e que, em dias de escola, se costumam deitar entre as 22 horas e a meia-noite. O que mais os diferencia, no conjunto das variáveis aqui consideradas, é o facto de, apesar de ambos os grupos terem práticas regulares de realização dos trabalhos de casa, os/as alunos/as do *cluster* 1 (condições mais favoráveis) parecerem ser mais cumpridores/as do que os/as do *cluster* 3, já que estes/as últimos/as estão fortemente associados/as à categoria «a maior parte das vezes». É ainda de referir que os/as alunos/as pertencentes ao *cluster* 3 (enquadramento misto e sucesso irregular) consideram satisfatório o seu desempenho escolar, ao passo que os/as jovens agrupados/as no *cluster* 1 o definem como «bom».

Quadro 33. Cruzamento da pertença aos *clusters* com outras variáveis de contexto e práticas escolares e de aprendizagem

Variável	Categoria		Cluster 1	Cluster 2	Cluster 3
Já alguma vez teve um processo disciplinar	Sim	Count	33	301	105
		Adjusted Residual	-5,8	10,9	-6,9
	Não	Count	183	343	401
		Adjusted Residual	5,8	-10,9	6,9
Tem quarto só para si	Sim	Count	185	377	392
		Adjusted Residual	5,5	-8,7	4,8
	Não	Count	31	269	114
		Adjusted Residual	-5,5	8,7	-4,8
Com que frequência faz os TPC?	Não tenho TPC	Count	20	508	54
		Adjusted Residual	-10,8	25,5	-18,3
	Sempre	Count	88	47	154
		Adjusted Residual	7,7	-11,9	6,5
	A maior parte das vezes	Count	86	37	249
		Adjusted Residual	4,5	-16,9	14,0
	Raramante	Count	21	38	44
		Adjusted Residual	1,3	-2,2	1,3
	Nunca	Count	1	16	5
		Adjusted Residual	-1,5	2,4	-1,4
Opinião acerca do desempenho escolar	Excelente	Count	7	17	9
		Adjusted Residual	0,9	0,5	-1,2
	Muito bom	Count	23	66	43
		Adjusted Residual	0,5	0,7	-1,1
	Bom	Count	106	203	167
		Adjusted Residual	4,8	-2,5	-1,1
	Satisfatório	Count	75	316	268
		Adjusted Residual	-4,3	0,5	2,7
	Insuficiente	Count	3	28	17
		Adjusted Residual	-1,8	1,6	-0,2
	Mau	Count	2	16	2
		Adjusted Residual	-0,7	3,0	-2,5
Em dias de escola a que horas se costuma deitar	Antes das 10 da noite	Count	31	117	105
		Adjusted Residual	-1,7	-0,3	1,6
	Depois das 10 da noite	Count	144	296	320
		Adjusted Residual	3,6	-6,9	4,4
	Depois da meia-noite	Count	28	125	43
		Adjusted Residual	-0,6	5,0	-4,7
	Não tenho hora para me deitar	Count	13	107	35
		Adjusted Residual	-2,7	5,8	-3,9
	Outra	Count	0	1	3
		Adjusted Residual	-0,9	-0,9	1,6

Quando não podemos estabelecer, à partida, o número de grupos a criar, deve ser usado um (ou vários) dos métodos de agrupamento hierárquico disponíveis. Para uma descrição desses métodos e exemplo de aplicação, ver Anexo 5.1.

É de notar que, no exemplo que agora apresentámos, a análise de *clusters* se constitui como uma operacionalização dos resultados obtidos a partir da ACM. Contudo, as suas potencialidades não se esgotam aqui e a sua aplicação é útil em todos os casos em que se queira agrupar elementos segundo a sua semelhança relativamente aos atributos de interesse.

Regressão Logística

A regressão logística constitui-se como um método de dependência que tem como objectivo a explicação e previsão de uma variável dependente (nominal dicotómica), em função do comportamento de *n* variáveis independentes (categóricas e/ou métricas). No âmbito da Sociologia, esta medida pode ser de grande utilidade, já que permite a avaliação do impacto simultâneo de um conjunto de atributos numa determinada variável. E, considerando que, nesta área, predominam as variáveis de tipo categórico, esta metodologia representa uma alternativa muito útil às regressões lineares.

A regressão logística enquadra-se no conjunto de métodos de regressão categorial que inclui também as regressões ordinais e multinomiais. O que as caracteriza é o facto de a variável dependente ser categórica (por oposição à regressão linear, onde a variável dependente é quantitativa): nominal dicotómica no caso da regressão logística; nominal policotómica no caso da regressão multinomial; e ordinal no caso da regressão ordinal. Podemos, então, recorrer à utilização de uma regressão logística para estimar a probabilidade de ocorrência (também designada como probabilidade de sucesso) de um determinado evento (variável dependente), dadas determinadas condições (variáveis independentes), sendo que as variáveis preditoras podem ser tanto quantitativas como categóricas).

A probabilidade de ocorrência da variável explicada (*Y*) é dada por $P(Y)=\frac{e^{(b_0+b_1X_1i+b_2X_{2i}+\ldots.+b_kX_{ki})}}{1+e^{(b_0+b_1X_1i+b_2X_{2i}+\ldots.+b_kX_{ki})}}$ (Equação 1), onde *P*(*Y*) é a probabilidade de *Y* ocorrer (probabilidade de sucesso), *e* corresponde ao logaritmo natural, b_0 representa a constante, X_1 a X_k correspondem às *k* variáveis explicativas e b_1 a b_k são os coeficientes associados às variáveis independentes (Field, 2013; Tabachnick & Fidell, 2013)[23]. Uma das grandes vantagens deste tipo de modelação é a de possibilitar a avaliação da magnitude da influência que as variáveis explicativas têm na variação da variável dependente.

Apenas para dar alguns exemplos no contexto dos problemas de investigação que são colocados na área da Sociologia, podería-

23 Em alternativa, a probabilidade de sucesso pode ser dada por $P(Y)=\frac{1}{1+e^{-(b_0+b_1X_1i+b_2X_{2i}+\ldots.+b_kX_{ki})}}$.

mos recorrer a uma regressão logística para estimar a probabilidade de uma pessoa votar nas eleições (por oposição a abster-se), mediante características sociodemográficas e do posicionamento político-ideológico; ou de uma mulher sofrer de ansiedade dadas as suas características físicas, psicológicas, sociodemográficas e de vitimação. Daremos, de seguida, um exemplo de aplicação de uma regressão logística, onde, recorrendo aos dados do inquérito aos custos sociais e económicos da violência contra as mulheres, realizado em 2002 (Lisboa *et al.*, 2006), pretendemos prever a probabilidade de uma mulher ser vítima de violência. Sendo que a variável dependente é dicotómica (0 – não vítima; 1 – vítima), a regressão logística é o instrumento adequado. Para esta análise, tomaremos como variáveis explicativas diversas características sociodemográficas das mulheres, sendo o modelo de partida constituído pelas seguintes variáveis (Quadro 34):

Quadro 34. Variáveis do modelo de regressão logística

Variável dependente	**Categorias**
Expectativas de escolaridade	Não vítima (R)
	Vítima
Variáveis independentes	**Categorias**
Idade	---
Estado civil	Solteira (R)
	Casada/união de facto
	Divorciada/separada
	Viúva
Nível de instrução	Não sabe ler e /ou escrever (R)
	1º ciclo
	2º ciclo
	3º ciclo/secundário
	Superior

Variável dependente	Categorias
Tem filhos	Não (R)
	Sim
Profissão	Quadros superiores da sdminstração pública, dirigentes e quadros superiores de empresa (R)
	Especialistas das profissões intelectuais e cientificas
	Técnicas e profissionais de nível intermédio
	Pessoal administrativo e similares
	Pessoal dos serviços e vendedoras
	Operárias, artífices e trabalhadoras similares; operadoras de instalações e máquinas e trabalhadoras da montagem
	Trabalhadoras não qualificadas
	Domésticas
	Estudantes
	Reformadas
	Desempregadas

(R) Categoria de referência

Este método apresenta a grande vantagem de não pressupor determinadas distribuições relativamente aos preditores (como no caso da regressão linear ou da análise discriminante, onde é necessária a normalidade das distribuições), o que não significa que não existam condições para a sua aplicação. Tal como descrevem Field (2013) e Tabachnick e Fidell (2013), é necessária a validação de alguns pressupostos de aplicação. Pela especificidade e extensão dos procedimentos, remetemo-los para anexo (Anexo 5.2).

Verificados os pressupostos, o passo seguinte consiste na definição do modelo, ou seja, na escolha das variáveis explicativas que faz sentido integrar na análise. Segundo Field (2013), devemos seleccionar o melhor modelo de regressão em blocos: no primeiro, incluímos apenas uma variável independente; no segundo, a mesma variável do primeiro bloco e uma nova; e por aí em diante, numa lógica cumulativa. Neste caso, porque estamos a considerar a introdução de cinco variáveis explicativas, vamos estimar um modelo com cinco blocos. Este procedimento permite-nos uma

comparação directa entre os vários modelos, já que, ao comparar os sucessivos ajustamentos, percebemos se a introdução de uma nova variável explicativa aumenta significativamente, ou não, a capacidade explicativa do modelo. Sabendo que a introdução de uma nova variável independente aumenta sempre a capacidade explicativa, queremos saber se esse aumento é significativo, ou seja, se aquilo que se ganha em explicação compensa a complexificação do modelo, tendo em vista a ideia da parcimónia: «The statistical implication of using a parsimony heuristic is that models be kept as simple as possible. In other words, do not include predictors unless they have explanatory benefit» (Field, 2013: 908).

Apresentamos, de seguida, os resultados das regressões logísticas em blocos (5 modelos em teste). Começamos pelo bloco de partida (Quadro 35), o modelo inicial, que é calculado antes da introdução de qualquer variável explicativa (onde apenas a constante entra no modelo), e que vai servir de base à comparação com o bloco seguinte.

Quadro 35. Selecção do modelo: Bloco 0 (excerto do *output*)

Block 0: Beginning Block

Iteration History[a,b,c]

Iteration		-2 Log likelihood	Coefficients
			Constant
Step 0	1	1821,391	-,819
	2	1820,575	-,869
	3	1820,575	-,870

a. is included in the model.
b. Initial -2 Log Likelihood: 1820,575
c. Estimation terminated at iteration number 3 because parameter estimates changed by less than ,001.

O bloco 1 (Quadro 36) resulta da introdução da «idade» como variável explicativa. Através do quadro *Omnibus tests of model coefficients*, percebemos que este modelo não é significa-

tivo ($\chi^2_{(1)} = 0{,}034$, $p = 0{,}854$), ou seja, a variável «idade» não contribui significativamente para a explicação da variável dependente.

O modelo 2 (Quadro 37) tem como variáveis explicativas a «idade» e o «estado civil». O modelo é significativo ($\chi^2_{(4)} = 26{,}284$, $p < 0{,}001$), ou seja, a combinação destas duas variáveis no modelo parece contribuir para a explicação da variável dependente. Para além disto, importa saber se a melhoria do ajustamento relativamente ao modelo (bloco) anterior é significativa; essa melhoria é dada pelo valor de χ^2 associado ao bloco. Este valor é o resultado da diferença entre o χ^2 do modelo 2 e do modelo 1 (com os graus de liberdade a serem calculados da mesma forma). Assim, a mudança no χ^2 é significativa ($\chi^2_{(3)} = 26{,}250$, $p < 0{,}001$), pelo que a introdução da nova variável tem efeito significativo no modelo.

Quadro 36. Selecção do modelo: Bloco 1 (testes globais aos coeficientes do modelo)

Block 1: Method = Enter				
Omnibus Tests of Model Coefficients				
		Chi-square	df	Sig.
Step 1	Step	,034	1	,854
	Block	,034	1	,854
	Model	,034	1	,854

Quadro 37. Selecção do modelo: Bloco 2 (testes globais aos coeficientes do modelo)

Block 2: Method = Enter				
Omnibus Tests of Model Coefficients				
		Chi-square	df	Sig.
Step 1	Step	26,250	3	,000
	Block	26,250	3	,000
	Model	26,284	4	,000

O modelo 3 (Quadro 38) resulta da introdução da «idade», do «estado civil» e do «ter filhos» como variáveis explicativas. O modelo é significativo ($\chi^2_{(5)} = 26{,}837$, $p < 0{,}001$). Contudo, a melhoria do ajustamento não é significativa ($\chi^2_{(1)} = 0{,}554$, $p = 0{,}457$), pelo que a introdução da nova variável não contribui para a melhoria do modelo.

O modelo 4 (Quadro 39) resulta da introdução da «idade», do «estado civil», do «ter filhos» e do «nível de instrução» como variáveis explicativas. O modelo é significativo ($\chi^2_{(9)} = 28{,}764$, $p = 0{,}001$). A melhoria do ajustamento volta a não ser significativa ($\chi^2_{(4)} = 1{,}927$, $p = 0{,}749$), pelo que a introdução da nova variável não contribui para a melhoria do modelo.

Quadro 38. Selecção do modelo: Bloco 3 (testes globais aos coeficientes do modelo)

Block 3: Method = Enter				
Omnibus Tests of Model Coefficients				
		Chi-square	df	Sig.
Step 1	Step	,554	1	,457
	Block	,554	1	,457
	Model	26,837	5	,000

Quadro 39. Selecção do modelo: Bloco 4 (testes globais aos coeficientes do modelo)

Block 4: Method = Enter				
Omnibus Tests of Model Coefficients				
		Chi-square	df	Sig.
Step 1	Step	1,927	4	,749
	Block	1,927	4	,749
	Model	28,764	9	,001

Por último, temos o modelo 5 (Quadro 40), que resulta da integração de todas as variáveis anteriores e da «profissão». O modelo é significativo ($\chi^2_{(19)} = 36{,}232$, $p = 0{,}010$). A melhoria do ajustamento não é significativa ($\chi^2_{(10)} = 7{,}468$, $p = 0{,}681$), pelo que a introdução da nova variável não tem um efeito significativo na melhoria do modelo relativamente ao anterior.

Quadro 40. Selecção do modelo: Bloco 5 (testes globais aos coeficientes do modelo)

Block 5: Method = Enter				
Omnibus Tests of Model Coefficients				
		Chi-square	df	Sig.
Step 1	Step	7,468	10	,681
	Block	7,468	10	,681
	Model	36,232	19	,010

Por forma a confirmar os resultados obtidos através desta metodologia, calculámos ainda o mesmo modelo, com estas cinco variáveis explicativas, através de um método *stepwise* (*forward LR*) que, ao invés de calcular os diferentes modelos em blocos, usa apenas um, mas estima o modelo em vários passos (que acabam por ser equivalentes aos blocos) e cujos resultados (que apresentamos no Anexo 5.3) corroboram as opções tomadas a partir das regressões em blocos. Nesse sentido, as evidências estatísticas

sugerem que optemos por um modelo com apenas uma variável explicativa. Contudo, em termos substantivos, não faz qualquer sentido que pretendamos explicar a vitimação das mulheres, fenómeno social complexo, apenas por um único atributo. Estes resultados indicam, portanto, e em linha com as conclusões retiradas em diversos estudos sobre violência contra as mulheres e violência doméstica, que não existe um perfil sociodemográfico comum às vítimas, e que a este tipo de violência é transversal ao tecido social.

Em todo o caso, e meramente como exemplo académico, completaremos a análise. Estimámos, então, o modelo final com as cinco variáveis explicativas (embora sabendo que apenas uma delas se revelará significativa), através do método *Enter* que, ao contrário dos métodos *stepwise*, não exclui variáveis do modelo com base na sua (não) significância. Passemos então à interpretação do modelo (Quadro 41). A forma mais comum e intuitiva de analisar os coeficientes de regressão é a de interpretar, não os betas (coeficientes de regressão), mas os exponenciais destes ($Exp(B)$), já que podem ser interpretados directamente sem ser necessária uma transformação logarítmica. Designados por *odds ratios*, representam a alteração da chance[24] de ocorrência de uma das categorias da variável dependente em resultado do aumento de uma unidade no preditor. Na prática, medem o aumento ou a diminuição da chance de ocorrência

24 É importante fazer a distinção entre *odds* e *probabilities* (que designamos por chances e probabilidades, respectivamente). Apesar de ambos os termos remeterem para a ideia de possibilidade de ocorrência, resultam de diferentes abordagens à possibilidade de um evento ocorrer, e representam termos que não são, portanto, intermutáveis. A probabilidade é dada pela divisão entre o resultado esperado e a totalidade de resultados possíveis (ex.: numa situação de moeda ao ar, a probabilidade de sair «coroa» é dada por 1/2, ou seja, um resultado esperado sobre dois resultados possíveis, resultando em 0,5, ou 50%). Os *odds* (ou chances) resultam da razão entre o número de resultados esperados e o número de resultados que não são os esperados; recorrendo ao mesmo exemplo da moeda ao ar, as chances de sair «coroa» são dadas por 1/1, ou seja, um resultado esperado (coroa) sobre um resultado contrário ao esperado (cara), resultando em 1:1. Diz-se, portanto, que as chances de sucesso, neste caso, são de um para um. Dito ainda de outra forma, se resolvermos a fracção (1/1), podemos dizer que as chances de sair «coroa» são de um (por cada «cara» que sair, deverá sair, em média, uma «coroa»). Ao passo que as probabilidades variam entre 0 e 1 (ou entre 0 e 100 se analisarmos em percentagem), as chances variam entre 0 e ∞. O *odds ratio* (ou rácio da chances) representa a razão entre duas chances, ou seja, é a chance de um acontecimento ocorrer tendo em conta as chances de ocorrência da outra categoria. Matematicamente, $OR = O_1/O_2 = \frac{p_1}{1-p_1} / \frac{p_2}{1-p_2}$.

da categoria de interesse[25]. *Odds ratios* superiores a 1 significam que o aumento de uma unidade no preditor provoca um aumento na chance de ocorrência da categoria de interesse (da variável dependente) e valores inferiores a 1 representam uma diminuição dessa mesma chance. Os *odds* são calculados pela divisão entre a probabilidade de o evento acontecer e a probabilidade de o evento não ocorrer e essa probabilidade é dada pela Equação 1.

Quadro 41. Modelo de regressão logística das características sociodemográficas como preditoras da probabilidade das mulheres serem vítimas

Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Idade	-,025	,160	,024	1	,876	,975	,713	1,334
	Solteira			25,413	3	,000			
	Casada	,028	,207	,018	1	,893	1,028	,685	1,544
	Divorciada	1,111	,276	16,175	1	,000	3,037	1,767	5,220
	Viúva	,298	,267	1,251	1	,263	1,347	,799	2,272
	Tem filhos	,098	,179	,297	1	,586	1,103	,776	1,567
	NS ler/escrever			2,118	4	,714			
	1º ciclo	-,200	,260	,591	1	,442	,819	,492	1,363
	2º ciclo	,048	,309	,024	1	,878	1,049	,572	1,921
	3º ciclo/secundário	-,026	,303	,007	1	,931	,974	,538	1,764
	Superior	-,104	,368	,079	1	,778	,901	,438	1,856
	Quadros			7,419	10	,685			
	Intelect	,225	,502	,201	1	,654	1,252	,469	3,347
	Técnica	,176	,462	,145	1	,704	1,192	,482	2,947
	Admin	,162	,506	,103	1	,749	1,176	,436	3,169
	Serviços	,343	,407	,711	1	,399	1,410	,634	3,132
	OperarioOpmaq	,017	,460	,001	1	,971	1,017	,413	2,504
	Trabnqualif	,439	,432	1,030	1	,310	1,550	,665	3,617
	Doméstica	,291	,425	,470	1	,493	1,338	,582	3,077
	Estududante	,541	,465	1,353	1	,245	1,717	,690	4,272
	Reformada	,363	,439	,683	1	,409	1,438	,608	3,401
	Desempregada	,799	,486	2,703	1	,100	2,222	,858	5,758
	Constant	-1,275	,571	4,988	1	,026	,280		

a. Variable(s) entered on step 1: Idade, Est_Civ, Filhos, N_Inst, Prof.

Nota: Quadros: Quadros Superiores da Administração Pública, Dirigentes e Quadros Superiores de Empresas; Intelect: Especialistas das Profissões Intelectuais e Científicas; Técnica: Técnicos e Profissionais de Nível Intermédio; Admin: Pessoal Administrativo e Similares; Serviços: Pessoal dos Serviços e Vendedores; OperarioOpmaq: Operários, Artífices e Trabalhadores Similares e Operadores de Instalações e Máquinas e Trabalhadores da Montagem; Trabnqualif: Trabalhadores não Qualificados (cf. Classificação Nacional de Profissões, versão 1994).

25 A categoria «vítima» foi codificada com o valor 1 e a categoria «não vítima» com o valor 0, pelo que esta é a categoria de referência, sendo a categoria de interesse a primeira, para a qual se retirarão as conclusões.

Os resultados obtidos revelam, como já tínhamos visto aquando da selecção do modelo, que a maior parte das características sociodemográficas não contribuem estatisticamente para explicar a variabilidade da vitimação. De facto, praticamente nenhuma das variáveis consideradas na análise (correspondendo aqui às categorias das variáveis seleccionadas) é significativa para o modelo (através da significância associada à estatística de Wald) (Quadro 41). Como se observa, apenas a categoria «divorciada» é significativa no modelo ($\chi^2_{(1)} = 16{,}175$, $p < 0{,}001$). Assim, e na sequência do que foi referido anteriormente, devemos interpretar o valor de *Exp(B)*: uma mulher divorciada tem uma chance três vezes superior (*Exp(B)* = 3,037) de ser vítima relativamente a uma mulher solteira (porque é esta a categoria de referência). Dito de outra forma, as chances de uma mulher divorciada ter sido vítima são 204% superiores às de uma mulher solteira ((3,037 – 1) * 100 = 203,7%). Voltamos a referir que este exercício foi feito a título meramente exemplificativo, já que este não pode ser considerado como um bom modelo explicativo e preditivo (ver Anexo 5.4) para avaliação do ajustamento do modelo seleccionado).

Bibliografia

Agresti, A. (2013). *Categorical data analysis*. New Jersey: John Wiley & Sons.

Allison, P. D. (2012). *Logistic regression using SAS: theory and application*. Cary: SAS Institute.

Benzécri, J.-P. (1976). *L'analyse des données*. (2 vols.). Paris: Dunod.

Bertolini, G.; D'amico, R.; Nardi, D.; Tinazzi, A. & Apolone, G. (2000). "One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model", in *Journal of Epidemiology and Biostatistics*, 5(4), pp. 251-253.

Carifio, J. & Perla, R. (2008), "Resolving the 50-year debate around using and misusing Likert scales", in *Medical Education, 42*, pp. 1150-1152.

Carvalho, H. (2008). *Análise multivariada de dados qualitativos – Utilização da ACM com o SPSS*. Lisboa: Sílabo.

COHEN, J.; COHEN, P.; WEST, S. & AIKEN, L. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. New Jersey: Lawrence Erlbaum Associates.

EVERITT, B. S.; LANDAU, S.; LEESE, M. & STAHL, D. (2011). *Cluster analysis*. West Sussex: Wiley.

FIELD, A. (2013). *Discovering statistics using IBM SPSS Statistics*. London: SAGE.

GARCIA PEREIRA, H. (2008). «Prefácio», in H. Carvalho (Ed.), *Análise multivariada de dados qualitativos – Utilização da ACM com o SPSS* (pp. 9-11). Lisboa: Sílabo.

HAIR, J. F. J.; HAIR, J. F.; BLACK, W. C.; BABIN, B. J. & ANDERSON, R. E. (2013). *Multivariate data analysis*. Essex: Pearson.

HÄRDLE, W. & SIMAR, L. (2007). *Applied multivariate statistical analysis*. Berlin: Springer.

HO, R. (2006). *Handbook of univariate and multivariate data analysis and interpretation with SPSS*. Boca Raton: CRC Press.

HOSMER, D. W.; LEMESHOW, S. & STURDIVANT, R. X. (2013). *Applied logistic regression*. New Jersey: Wiley.

IBM (2014). *Multicollinearity diagnostics for Logistic Regression, NOMREG, or PLUM*. Acedido a 28/03/2016 em http://www-01.ibm.com/support/docview.wss?uid=swg21476696.

JAIN, A. K.; MURTY, M. N. & FLYNN, P. J. (1999). "Data clustering: a review", in *ACM computing surveys*, *31* (3), pp. 264-323.

KRAMER, A. A. & ZIMMERMAN, J. E. (2007). "Assessing the calibration of mortality benchmarks in critical care: the Hosmer-Lemeshow test revisited", in *Critical Care Medicine*, *35* (9), pp. 2052-2056.

LAUREANO, R. (2013). *Testes de hipóteses com o SPSS – O meu manual de consulta rápida*. Lisboa: Sílabo.

LILLIEFORS, H. W. (1967). "On the Kolmogorov-Smirnov test for normality with mean and variance unknown", in *Journal of the American Statistical Association*, *62* (318), pp. 399-402.

LISBOA, M. (2014). *A importância das metodologias de investigação na construção da Sociologia como ciência: o refinamento das metodologias quantitativas*. Lição de Agregação em Sociologia, ramo Teorias e Metodologias, no âmbito da disciplina de Análise de Dados Multivariada. FCSH-UNL.

LISBOA, M.; CARMO, I. d.; VICENTE, L. B.; NÓVOA, A.; BARROS, P. P.; ROQUE, A.; SILVA, S. M.; FRANCO, L. & AMÂNDIO, S. (2006). *Prevenir ou remediar – Os custos sociais e económicos da violência contra as mulheres*. Lisboa: Colibri.

Lisboa, M., & Malta, J. (2009). «Infâncias adiadas: análise dos contextos sociais, económicos e culturais favoráveis à produção e reprodução do trabalho infantil», in M. Lisboa (Ed.), *Infância interrompida. Caracterização das actividades desenvolvidas pelas crianças e jovens em Portugal* (pp. 83-113). Lisboa: Colibri; PETI; CESNOVA; SociNova.

Lisboa, M.; Sarmento, M. J.; Justino, D.; Valente Rosa, M. J.; Malta, J.; Carvalho, M. J. L.; Leandro, A.; Pinho, P.; Graça, E. & Fonte, E. (2009), *Infância interrompida. Caracterização das actividades desenvolvidas pelas crianças e jovens em Portugal*. Lisboa: Colibri; PETI; CESNOVA; SociNova.

Lourenço, N.; Lisboa, M. & Pais, E. (1997). *Violência contra as mulheres*. Lisboa: Comissão para a Igualdade e para os Direitos das Mulheres.

Marôco, J. (2014). *Análise Estatística com o SPSS Statistics*. Lisboa: Report Number.

Pampel, F. C. (2000). *Logistic Regression: A Primer*. Thousand Oaks: SAGE.

Reis, E. (2001). *Estatística multivariada aplicada*. Lisboa: Sílabo.

Reis, E. (2009). *Estatística descritiva*. Lisboa: Sílabo.

Reis, E.; Melo, P.; Andrade, R. & Calapez, T. (2016). *Estatística aplicada, volume 2*. Lisboa: Sílabo.

Sharpe, D. (2015). "Your chi-square test is statistically significant: now what? Practical Assessment", in *Research & Evaluation*, *20* (8), pp. 1-10.

Spicer, J. (2005). *Making sense of multivariate data analysis: an intuitive approach*. Thousand Oaks: SAGE.

Tabachnick, B. G. & Fidell, L. S. (2013). *Using multivariate statistics*. Boston: Pearson.